

# DYNAMICS OF SOCIAL NETWORK EVOLUTION AND INFORMATION DIFFUSION

A Dissertation

Presented to the Faculty of the Graduate School  
of Cornell University

in Partial Fulfillment of the Requirements for the Degree of  
Doctor of Philosophy

by

Daniel Mauricio Romero

May 2012

© 2012 Daniel Mauricio Romero  
ALL RIGHTS RESERVED

# DYNAMICS OF SOCIAL NETWORK EVOLUTION AND INFORMATION DIFFUSION

Daniel Mauricio Romero, Ph.D.

Cornell University 2012

Millions of interactions between people take place on the Web everyday. In this work, we utilize data obtained by tracking these interactions on social media sites to study two important aspects of social networks: the way in which connections between people form and evolve over time, and the dynamics of information diffusion on the network. We introduce novel methodologies, algorithms, and mathematical models to analyze observations from rich datasets. Our results validate known sociological theories of link formation and information diffusion at large scale and suggest new ones.

We study the formation of links in the network of interactions among people in social media sites from the premise that these networks are inherently different from offline social networks. Online interaction networks are not purely social, but a combination of social and information networks. We introduce mechanisms of link formation that are motivated from sociological theories of social network formation, and are generalized to social-information networks. Furthermore, we study how the communication patterns of connected users of social media sites change as a response to new connections arriving to the network and compare our results to the predictions that various sociological theories would suggest.

There is an intuitive sense in which the network of interactions among people is related to how information spreads on the network. In this work, we

show that this relationship is present in both directions. That is, the structure of the network can determine whether information spreads through the network, and the kind of information users are exposed to can determine the connections among the users. Furthermore, we show that the dynamics of information diffusion can change significantly depending on the topic, which suggests the mechanisms that control information diffusion are context dependent.

## **BIOGRAPHICAL SKETCH**

Daniel M. Romero was born in Bogota, Colombia on May 24th, 1983 and moved to Scottsdale, Arizona at age 15. He became interested in Mathematics during his high school years and obtained a Bachelor's and a Master's degree in Mathematics at Arizona State University. In 2007, he moved to Ithaca, New York to work on a Doctorate degree at Cornell University's Center for Applied Mathematics. During the early years of his Ph.D, he became interested in mathematical applications to research questions related to Social Sciences. He is expected to finish his Ph.D in May of 2012.

To my wife, my daughter, my parents, my brother, and my grandparents.

## ACKNOWLEDGEMENTS

First of all, I would like to thank Jon Kleinberg for all of his support throughout these years. He has been the best advisor anyone could ever hope to have. I have learned many invaluable lessons from him. His guidance has been an essential part of my success in not only completing my Ph.D but also in growing as a researcher and a professional. All of my accomplishments during my doctoral studies would have been impossible without his mentoring. It is amazing how a short conversation with Jon can provide so much insight into the direction of a research project. Jon is truly an amazing investigator, mentor, and teacher, and one of the nicest persons I have ever met.

I was fortunate to spend six months in 2008 and a summer in 2010 working in HP's Social Computing Lab, where I had the pleasure of meeting Bernardo Huberman. I was fortunate to collaborate with him and some of his team members on several research papers. Bernardo has had an important influence in shaping my research interests and research agenda. He has taught me the importance of looking at the larger context of the research questions I aim to answer.

When I arrived at Cornell, I was lucky to have Steven Strogatz as my mentor. He provided me with important guidance when I started to become interested in studying social networks. To this day, he continues to be very supportive and providing me with invaluable professional and scientific mentoring.

I had the pleasure of spending two summers in Microsoft Research where I met very smart and interesting people. I would particularly like to thank Christian Borgs, Jennifer Chayes, Nina Mishra, and Panayiotis Tsaparas for their great mentoring during my internships, which made my time in Microsoft very enjoyable and productive.

Throughout my doctoral studies I have had the pleasure of collaborating with very bright people who have made my research possible. Many of them have become my good friends. I would like to thank everyone of them: Ram Asur, Vlad Barash, Kamal Barley, Christian Borgs, danah boyd, Mike Brzozowski, Carlos Castillo-Chavez, Jennifer Chayes, Justin Cheng, Wojtek Galuba, Bernardo Huberman, Dan Huttenlocher, Jon Kleinberg, Christopher Kribs-Zaleta, Brendan Meeder, Nina Mishra, Anuj Mubayi, Clara Orbe, Grant Schoenebeck, Chenhao Tan, Panayiotis Tsaparas, Johan Ugander, Alicia Urdapilleta, Alex Vladimirsky, Fang Wu, and Sarita Yardi.

I would like to thank the members of my committee, Jon Kleinberg, Steven Strogatz, John Hopcroft, and Dan Huttenlocher for all their support and useful comments about my research.

Before coming to Cornell, I had the pleasure of meeting Carlos Castillo-Chavez in Arizona State University. Thanks to Carlos and his research program, I was introduced to the world of scientific research. He played a very influential role in my decision to pursue a Ph.D and to become a scientist. I thank him for his dedication in making a big difference in my life and so many talented students' lives. He is a true role model.

I would like to thank my wife Alicia Urdapilleta. Her love, support, and encouragement have been the most important ingredients for my success in completing my Ph.D. She has been an amazing role model and companion during this journey and I am extremely lucky to have her by my side. Thank you for being there for me throughout these years. There is absolutely no way I could have done this without you. My daughter Ylani has also been a great source of inspiration for me. Even though she does not know it yet, she gives me the energy and motivation to work hard every day.



Finally, I would like to thank my parents, Yolanda Camacho and Gerson Romero, and my brother Diego Romero for their unconditional love and support. I thank my parents for instilling in me the importance of education and hard work. Thanks to their own hard work and dedication I had the opportunity to get an education and live my dreams. I owe all of my accomplishments to them.

## TABLE OF CONTENTS

Biographical Sketch . . . . .	iii
Dedication . . . . .	iv
Acknowledgements . . . . .	v
Table of Contents . . . . .	viii
List of Tables . . . . .	xi
List of Figures . . . . .	xii
<b>1 Introduction</b>	<b>1</b>
1.1 Overview . . . . .	1
1.2 Topics . . . . .	3
1.2.1 Social and Informational Ties . . . . .	3
1.2.2 Link Formation in a Social-Information Network . . . . .	4
1.2.3 Maintaining Social Ties . . . . .	5
1.2.4 Information Flow Across Different Topics . . . . .	5
1.2.5 Influence and Passivity in Social Networks . . . . .	7
1.2.6 Interplay between Network Structure and Information Flow . . . . .	8
<b>2 Literature Review</b>	<b>10</b>
2.1 Generative Models for Graph Formation . . . . .	10
2.1.1 Small-World Networks . . . . .	10
2.1.2 Preferential Attachment . . . . .	13
2.2 Beyond Static Simple Graphs . . . . .	14
2.2.1 Weighted Graphs . . . . .	14
2.2.2 Signed Graphs . . . . .	16
2.2.3 Evolving Graphs . . . . .	18
2.3 Information Diffusion . . . . .	20
2.3.1 Basic Models of Information Diffusion . . . . .	20
2.3.2 Online Diffusion Processes . . . . .	22
2.3.3 Influence . . . . .	23
<b>I Network Structure Evolution</b>	<b>25</b>
<b>3 Social-Information Networks</b>	<b>26</b>
3.1 Results . . . . .	27
3.2 Discussion . . . . .	32
<b>4 The Directed Closure Process in Social-Information Networks</b>	<b>36</b>
4.1 Twitter Data and Micro-Celebrities . . . . .	42
4.2 Evidence for Directed Closure . . . . .	44
4.3 Preferential attachment . . . . .	46
4.4 Preferential Attachment with Fitness . . . . .	49
4.5 Preferential Attachment with Communities . . . . .	53

4.6	Discussion . . . . .	55
<b>5</b>	<b>Maintaining Ties in Social Media</b>	<b>58</b>
5.1	Data Set and Network Extraction . . . . .	64
5.2	Balance Vs. Betweenness . . . . .	66
5.3	Exchange Theory and Spill-Over Effects . . . . .	69
5.4	Basic Properties of Relationship Decay . . . . .	73
5.5	Discussion . . . . .	78
<b>II</b>	<b>Information Difussion</b>	<b>81</b>
<b>6</b>	<b>The Mechanics of Information Flow by Topic</b>	<b>82</b>
6.1	Introduction . . . . .	82
6.2	Dataset, Network Definition, and Hashtag Classification . . . . .	90
6.3	Exposure Curves . . . . .	93
6.4	The structure of initial sets . . . . .	102
6.5	Simulations . . . . .	104
6.5.1	The Simulated Model . . . . .	105
6.5.2	Simulation Results . . . . .	106
6.6	Discussion . . . . .	109
<b>7</b>	<b>Influence and Passivity in Social Media</b>	<b>112</b>
7.1	Related work . . . . .	114
7.2	Graph Construction . . . . .	115
7.3	The IP Algorithm . . . . .	116
7.4	Evaluation . . . . .	121
7.4.1	Computations . . . . .	121
7.4.2	Influence as a correlate of attention . . . . .	122
7.5	IP Algorithm Adaptability . . . . .	128
7.6	Case Studies . . . . .	130
7.7	Discussion . . . . .	133
<b>III</b>	<b>Interplay between Network Evolution and Information Difussion</b>	<b>136</b>
<b>8</b>	<b>Social-Topical Affiliations</b>	<b>137</b>
8.1	Dataset . . . . .	139
8.2	Topical features predict links . . . . .	140
8.2.1	Measuring topic distance . . . . .	141
8.2.2	Predictive model . . . . .	143
8.2.3	Predictive model with edges . . . . .	145
8.2.4	Routing . . . . .	147

8.3	Social Adoption of Hashtags and Future Users . . . . .	149
8.3.1	Correlations between initial graph structure and popularity	150
8.3.2	Exogenous forces vs. Virality . . . . .	152
8.3.3	Predicting growth from structure . . . . .	155
8.3.4	Longer prediction horizons . . . . .	158
8.4	Related Work . . . . .	161
8.5	Discussion . . . . .	161
<b>9</b>	<b>Conclusion</b>	<b>164</b>
	<b>Bibliography</b>	<b>169</b>

## LIST OF TABLES

6.1	Definitions of categories used for annotation. . . . .	91
6.2	A small set of examples of members in each category. . . . .	91
6.3	Median values for number of mentions, number of users, and number of mentions per user for different types of hashtags . . .	102
6.4	Comparison of graphs induced by the first 500 early adopters of political hashtags and average hashtags. Column definitions: I. Average degree, II. Average triangle count, III. Average entering degree of the nodes in the border of the graphs, IV. Average num- ber of nodes in the border of the graphs. The error bars indicate the 95% confidence interval of the average value of a randomly selected set of hashtags of the same size as Political. . . . .	104
7.1	Users with the most IP-influence (with at least 10 URLs posted in the period) . . . . .	130
7.2	Users with the most IP-passivity . . . . .	131
7.3	Users with many followers and low relative influence . . . . .	132
7.4	Users with very few followers but high relative influence . . . .	133
8.1	Prediction accuracies for directed and mutual edges, as trained on the full set of hashtag features, individual hashtag features, and edge features. Accuracy was evaluated using 10-fold cross- validation on a balanced classification dataset. . . . .	144
8.2	Accuracy of a logistic regression mode for predicting whether a hashtag will double the number of adopters at different starting points: the 1000, 2000, and 4000 initial adopters, for both the follower and the @-message graphs. The accuracy of all models was evaluated using 10-fold cross-validation. . . . .	154

## LIST OF FIGURES

3.1	Number of posts as a function of the number of followers. The number of posts initially increases as the number of followers increases but it eventually saturates. . . . .	29
3.2	Number of posts as a function of the number of friends. The number of posts increases as the number of friends increases, reaching 3200 without saturating. . . . .	30
3.3	Histogram of contributor's number of friends divided by the number of followees. Most users have a very small number of friends compared to the number of followees they declared. . . .	31
3.4	Number of friends as a function of the number of followees. The total number of friends saturates while the number of followees keeps growing due to the minimal effort required to add a followee. . . . .	33
3.5	Proportion of friends vs. followees as a function of followers. It initially increases but rapidly approaches zero as the number of followees increases. . . . .	34
4.1	(a) Triadic closure in an undirected graph produces a triangle when an edge connects two nodes who already have a common neighbor. (b) Analogously, in a directed information network, directed closure occurs when a node $A$ links to a node $C$ to which it already has a two-step path (through a node $B$ ). This creates a directed triangle (a "feed-forward" structure on three nodes). . .	37
4.2	In this example, the edge from $A$ to $C$ exhibits closure if there is already a two-step path from $A$ to $C$ (i.e., through $B_1, B_2, B_3$ ) when the $A$ - $C$ edge arrives. . . . .	38
4.3	Closure ratio as a function of the arrival order of incoming edges for 18 Twitter $\mu$ -celebrities. The following are the professions of the $\mu$ -celebrities in each figure (from top to bottom curve). Top figure: Journalist, Venture Capital Blogger, Actor, Actor, DJ, Skateboarder. Middle figure: Comedian, Film Producer, Social Media Blogger, Musician, Actor, Journalist. Bottom figure: Comedian, TV Presenter, Actor, Musician, Filmmaker, Actor. . . . .	43
4.4	The connected dots indicate the actual value of $f_k$ , the circles indicate the average closure ratio among the 100 simulations, and the plus signs indicate the error bars. Results for 3 $\mu$ -celebrities are shown. The trend is similar for all other $\mu$ -celebrities . . . . .	45
4.5	Results from the preferential attachment simulation with $N = 200,000$ , $\alpha = .3$ , and $D = 10$ . The figure shows the closure ratio as a function of edge arrival order of the 10 nodes with highest in-degree. . . . .	47

4.6	The actual closure ratio of each node $j$ generated by the preferential attachment model with parameters $N = 200,000$ , $\alpha = .3$ , and $D = 10$ (dots) and its approximation by $C_{N-1}(j)$ (plus signs).	50
4.7	Results from the preferential attachment with fitness simulation with $N = 200,000$ , $\alpha = .3$ , and $D = 10$ . The top figure shows the closure ratio as a function of in-degree of the 10 nodes with highest in-degree. The bottom function shows the final closure ratio of each node $j$ (dots) and its approximation by $C_{N-1}(j)$ (plus signs).	51
4.8	Closure ratio as a function of In-Degree.	53
4.9	Closure ratio as a function of the Sum of In-Degree of Incoming Nodes.	54
4.10	The closure ratio as a function of in-degree for the 10 nodes with highest in-degree. Preferential attachment with communities simulation with $N = 200,000$ , $\alpha = .3$ , $\beta = .8$ , $C = 1,000$ , and $D = 10$ .	56
4.11	Results from the preferential attachment with communities simulation with $N = 200,000$ , $\alpha = .3$ , $\beta = .8$ , $C = 1000$ , and $D = 10$ .	57
5.1	The theories of balance and exchange postulate the effect of A and C forming a relationship on the B-A and B-C relationships.	59
5.2	Outside influence: The A-B relationship is potentially weakened not only by additional relationships within the online social network, but also by activities that altogether draw users away from the network.	61
5.3	Betweenness postulates that A is more dependent on B for information when A connects to nodes that are not connected to B than when she connects to nodes connected to B.	62
5.4	Percentage of message from A to B vs. the number of day after creation of open triad. The green curve is based on the $d$ -open triads and the red curve is based on the $d$ -closed triads. A must have sent from 200 to 1000 messages in total after day = 0 and A must have sent at least one messages on days 1, $d$ , and $2d$ .	67
5.5	Number of messages A sends to everyone but B vs. number of messages A sends to B, 3 days after the creation of the A-B edge.	70
5.6	Percentage of messages that A sends to B as a function of the percentage of A's non-B messages that go to friends of B. These messages take place 3 days after the creation of the A-B edge.	71
5.7	Number of messages A sends to friends of B vs. number of messages A sends to B, five days after the creation of the A-B edge. Node A sent exactly 10 messages in total to users other than B.	72
5.8	Zoom-in of figure 5.4(d). We observe jumps on the green curve at days $d$ and $2d$ and on the red curve at day $2d$ but not on day $d$ .	74

5.9	Probability that A will send a message to B $d$ days after having sent her one . . . . .	75
5.10	Average function $M$ for pairs $(A, B)$ in which A sent B at least 100 @-messages . . . . .	76
5.11	Average of $\log(M(n))$ as a function of $\log(n)$ where $n > 0$ (in blue) and as a function of $\log(-n)$ for $n < 0$ (in red) . . . . .	77
5.12	Average of $\log(M(n))$ as a function of $\log(n)$ where $n > 0$ (in blue for unreciprocated links and green for reciprocated ones) and as a function of $\log(-n)$ for $n < 0$ (in red for unreciprocated links and black for reciprocated ones) . . . . .	79
6.1	Average exposure curve for the top 500 hashtags. $P(K)$ is the fraction of users who adopt the hashtag directly after their $k^{th}$ exposure to it, given that they had not yet adopted it . . . . .	85
6.2	$F(P)$ for the different types of hashtags. The black dots are the average $F(P)$ among all hashtags, the red x is the average for the specific category, and the green dots indicate the 90% expected interval where the average for the specific set of hashtags would be if the set was chosen at random. Each point is the average of a set of at least 10 hashtags . . . . .	95
6.3	Sample exposure curves for hashtags #cantlivewithout (blue) and #hcr (red). . . . .	96
6.4	Point-wise average influence curves. The blue line is the average of all the influence curves, the red line is the average for the set of hashtags of the particular topic, and the green lines indicate the interval where the red line is expected to be if the hashtags were chosen at random. . . . .	97
6.5	Example of the approximation of an influence curve. The red curve is the influence curve for the hashtag #pickone, the green curves indicate the 95% binomial confidence interval, and the blue curve is the approximation. . . . .	100
6.6	Validating Category Differences: The median cascade sizes for three different categories. In (a) we randomize over the $p(k)$ curves and show that celebrity $p(k)$ curves don't perform as well as random $p(k)$ curves on celebrity start sets. Figures (b) and (c) illustrate the strength of the starting sets for political and idiom hashtags compared to random start sets. All starting sets consist of 500 users. . . . .	107



7.1	Evidence for the Twitter user passivity. We measure passivity by two metrics: 1. the user retweeting rate and 2. the audience retweeting rate. The <i>user retweeting rate</i> is the ratio between the number of URLs that user $i$ decides to retweet to the total number of URLs user $i$ received from the followed users. The <i>audience retweeting rate</i> is the ratio between the number of user $i$ 's URLs that were retweeted by $i$ 's followers to the number of times a follower of $i$ received a URL from $i$ . . . . .	117
7.2	IP-algorithm convergence. In each iteration we measure the sum of all the absolute changes of the computed influence and passivity values since the previous iteration . . . . .	122
7.3	We consider several user attributes: the number of followers, the number of times a user has been retweeted, the user's PageRank, H-index and IP-influence. For each of the 3.2M Bit.ly URLs we compute the average value of a user's attribute among all the users that mentioned that URL. This value becomes the $x$ coordinate of the URL-point; the $y$ coordinate is the number of clicks on the Bit.ly URL. The density of the URL-points is then plotted for each of the four user attributes. The solid line in each figure represents the 99.9th percentile of Bit.ly clicks at a given attribute value. The dotted line is the linear regression fit for the solid line with the fit's $R^2$ and slope displayed beside it. . . . .	123
7.4	For each user we place a user-point with IP-influence as the $y$ coordinate and the $x$ coordinate set to the number of user's followers. The density of user-points is represented in grayscale. The correlation between IP-influence and #followers is 0.44. . . .	127
7.5	The correlation between the IP-influence values computed based on two inputs: the co-mention influence graph and the retweet influence graph. The correlation between the two influence values is 0.06. . . . .	128
8.1	Edge density heterogeneity for the 100 most common hashtags in the dataset. . . . .	146
8.2	Linkage probability as a function of smallest common hashtag. (a) The probability of a given user following another user as a function of the size, and (b) the probability of a given user @-messaging another user as a function of size. Both figures are log-log scale. . . . .	147

8.3	Median number of final adopters as a function of the number of (a) edges and (b) singletons in the graph induced by the 1000 initial adopters, using a sliding window. Probability that hashtags will exceed $k$ adopters given the number of (c) edges and (d) singletons in the graph induced by the 1000 initial adopters, using a sliding window. From top to bottom, $K = 1500, 1750, 2000, 2500, 3000, 3500, 4000$ . We observe that hashtags with many or few singletons and edges are more likely to grow than hashtags with intermediate amounts. . . . .	150
8.4	Distribution of the structural features of the subgraphs induced by the 1000, 2000, and 4000 initial adopters. We see that while the edge count exhibits a heavy tailed distribution, the number of singletons, components and the size of the largest component are all broadly distributed over their support. . . . .	152
8.5	Prediction accuracy, precision, recall, and F1 score when predicting whether a hashtag will exceed a certain size using our logistic regression model based on graph structure. Models were trained using 5-fold cross validation, applied to those 7397 hashtags that reached a size of 1000. . . . .	159

# CHAPTER 1

## INTRODUCTION

### 1.1 Overview

The advent of the Web and online social networks have provided people all over the world with a tool to maintain communication with each other and to stay informed. On the research side, the online interactions among people generate very large and detailed data sets which provide an opportunity to study social phenomena and human behavior at a scale that was unimaginable in the past. Sociologists have researched the formation of social networks and the interactions among people for a long time. We can now complement the research methods of sociology with approaches that draw on large data sets and new algorithms. We can validate, at large scale, many of the theories they have developed and discover new ones. We can also study phenomena that cannot be observed without massive data sets, the kind of patterns that emerge at large scale but are invisible at the local level.

Before the Web, very few people were able to communicate information to the world at large. With the arrival of online social networks, blogging, and micro-blogging sites, information can come from anyone and reach millions of people very fast. We no longer depend on a very small number of people to learn what happens around the globe. Hence, the dynamics in which we communicate have drastically changed in the past few years. We are able to track much of the information that flows through the Web. In particular, we are able to track it on sites such as Twitter, Facebook, Google Plus, and Wikipedia, which enables us to learn about the way in which information spreads through

a network. For example, we are able to begin to answer questions such as: what determines whether a certain piece of information will become popular? Understanding the way in which information flows online is important for developing tools that allow people to communicate more efficiently. In this direction, people have studied the typical patterns by which information spreads[3, 8, 26, 28, 49, 80, 86, 89, 120]. In this work, we complement their findings by discovering ways in which these patterns vary across different topics and using these traces of information flow we propose novel algorithms to detect influential nodes in the network.

People started using online social networks in order to keep in touch with their friends, relatives, acquaintances, etc. Now the use of these networks is much broader; people now also use online social networks in order to stay informed about what is going on around the world. For example, CNN has a Facebook page where they periodically update the current breaking news and people can comment on them, share them with their friends, or simply stay informed. This means that the graphs of these online social networks are a combination of friendship edges and informational edges those edges that get formed because a user is interested in the information he can receive from the other user, not because of a personal relationship. In light of this phenomenon, it is important to understand the dynamics that govern the formation of networks that are not purely social or informational but a combination of both. In this work, we propose a mechanism of link formation that generalizes current theories of social network formation to social-information networks. We also conduct empirical studies that analyze how the structure of these networks changes over-time and how communication among connected nodes changes after new links arrive.

## **1.2 Topics**

Throughout this work, we use traces of how people interact and connect online to study various aspects of social behavior and communication. In particular, we use Twitter data to investigate two major topics: How social networks evolve over time and how information flows through them. Part I of the thesis deals of different aspects of network evolution and part II deals with information diffusion. Finally, while these two broad areas can be studied separately by assuming a fixed social network and analyzing information spread through its links, and by analyzing how links between nodes appear and change over time without considering information flow, in part III we will see that network evolution and information flow actually affect each other and one can study the interplay between the two.

### **1.2.1 Social and Informational Ties**

One of the reasons people use social media sites is to maintain social relationships online. However, they also use social media sites to maintain other kinds of relationships. Scholars, advertisers, and political activists see massive social media networks as a representation of social interactions that can be used to study the propagation of ideas, social bond dynamics and viral marketing. Hence, they are present in the network and also connect with regular users. For example, many celebrities, brands, politicians, businesses, and organizations in general have Twitter accounts and are followed by millions of people. This benefits the regular users because they stay updated with relevant information and it also benefits the organizations because they are able to reach the people who

have an interest in them. Therefore, there are two kinds of relationships in the network: the social ties and those that connect people to the organizations they are interested in. We call the latter informational ties. In chapter 3 we discuss how scarcity of attention and the daily rhythms of life and work makes people default to interacting with those few that matter to them and that reciprocate their attention. These are often the strong social ties of the user and not the informational ties. We present a study of social interactions within Twitter that reveals that the driver of usage is a sparse and hidden network of strong connections underlying the declared set of friends and followers.

### **1.2.2 Link Formation in a Social-Information Network**

Social media data are interesting to study because they provide a way to investigate theories from sociology about how we connect with each other. One of those well-known theories is triadic closure which states that when two people have a friend in common, they are more likely to establish a relationship than if they did not have a friend in common [107]. This fundamental theory of link formation has been empirically studied for pure social networks [71], but it has not been looked at empirically for directed networks such as Twitter where some of the ties are social and some are informational. In Chapter 4, we develop a formalization and methodology for studying a generalized version of directed closure which we call the *directed closure process*. We provide evidence for its important role in the formation of links on Twitter. We then analyze a sequence of models designed to capture the structural phenomena related to directed closure that we observe in Twitter data.

### **1.2.3 Maintaining Social Ties**

Part of understanding how a social network evolves is understanding how new links affect existing relationships. When users interact with one another on social media sites, the volume and frequency of their communication can shift over time, as their interaction is affected by a number of factors. In particular, if two users develop mutual relationships to third parties, this can exert a complex effect on the level of interaction between the two users—it has the potential to strengthen their relationship, through processes related to triadic closure, but it can also weaken their relationship, by drawing their communication away from one another and toward these newly formed connections. In chapter 5, we analyze the interplay of these competing forces and relate the underlying issues to classical theories in sociology—the theory of balance, the theory of exchange, and betweenness. Our setting forms an intriguing testing ground for these two theories, in that it provides a scenario in which their qualitative predictions are largely at odds with one another. In the course of our analysis, we also provide novel approaches for dealing with a common methodological problem in studying ties on social media sites: the tremendous volatility of these ties over time makes it hard to compare one's results to simple baselines that assume static or stable ties, and hence we must develop a set of more complex baselines that takes this temporal behavior into account.

### **1.2.4 Information Flow Across Different Topics**

The main purpose that online social networks serve its users is the ability to communicate and share information with other users they are linked to. While

it is possible for users to communicate with others that they are not connected to, most online social networks are designed in a way that makes users more likely to share information with their neighbors than with others. Hence, understanding how the structure of these networks changes is important because it controls the way information can flow on the network. A number of studies about on-line information diffusion have emerged in the past few years [3, 8, 26, 28, 49, 80, 86, 89, 120]. Much of this research has been focused on understanding the general mechanics of information diffusion that hold for any type of information. However, there is a widespread intuitive sense that different kinds of information spread differently on-line. It has been difficult to evaluate this question quantitatively since it requires a setting where many different kinds of information spread in a shared environment. In chapter 6, we study this issue on Twitter, analyzing the ways in which tokens known as hashtags spread on a network defined by the interactions among Twitter users. We find significant variation in the ways that widely-used hashtags on different topics spread. Our results show that this variation is not attributable simply to differences in stickiness, the probability of adoption based on one or more exposures, but also to a quantity that could be viewed as a kind of persistence—the relative extent to which repeated exposures to a hashtag continue to have significant marginal effects. We find that hashtags on politically controversial topics are particularly persistent, with repeated exposures continuing to have unusually large marginal effects on adoption; this provides, to our knowledge, the first large-scale validation of the complex contagion principle from sociology, which posits that repeated exposures to an idea are particularly crucial when the idea is in some way controversial or contentious. Among other findings, we discover that hashtags representing the natural analogues of Twitter idioms and



neologisms are particularly non-persistent, with the effect of multiple exposures decaying rapidly relative to the first exposure. We also study the subgraph structure of the initial adopters for different widely-adopted hashtags, again finding structural differences across topics. We develop simulation-based and generative models to analyze how the adoption dynamics interact with the network structure of the early adopters on which a hashtag spreads.

### **1.2.5 Influence and Passivity in Social Networks**

While social media sites provide users a venue where they can express their ideas and thoughts to potentially the entire world through viral spread on the network, it is not the case that every person has an equal chance of getting their ideas to become popular. Indeed, those who are already popular in the offline world, tend to be the ones who have the most connections online and hence have a better chance of reaching more people. The way in which social media is different from other kinds of media is that when someone attempts to spread their message, they rely on others to pass it on for them. In chapter 7, we conduct a study of information propagation within Twitter which reveals that the majority of users act as passive information consumers and do not forward the content to the network. Therefore, in order for individuals to become influential they must not only obtain attention and thus be popular, but also overcome user passivity. By studying information diffusion we are able to identify who the most influential users in the network are. We propose an algorithm that determines the influence and passivity of users based on their information forwarding activity. An evaluation performed with a 2.5 million user dataset shows that our influence measure is a good predictor of URL clicks, outper-

forming several other measures that do not explicitly take user passivity into account. We demonstrate that high popularity does not necessarily imply high influence and vice-versa.

## **1.2.6 Interplay between Network Structure and Information Flow**

As we observe in chapter 6, a user's connections can impact the information she is willing to share. At the same time, as we observe in chapter 4, users may decide who they will follow in the future based on new information they get from the current connections. For example, users may decide to follow a celebrity on Twitter after finding about it through their friends. This suggests that there is a fundamental relationship between information diffusion and network structure. However, measuring the extent to which they affect each other has remained an open question. In chapter 8, we examine the interface of two decisive structures forming the backbone of online social media: the graph structure of social networks and the set structure of topical affiliations who talks about what. In studying this interface, we identify key relationships whereby each of these structures can be understood in terms of the other. We look at the Twitter social network of both follower relationships and communication relationships, alongside the affiliations outlined by the hashtags used by people to label their communications. On Twitter, we demonstrate how the hashtags that a user adopts can be used to predict their social relationships, and also how the social relationships between the adopters of a hashtag can be used to predict the future popularity of that hashtag. We find that both relationships are driven by highly

computationally simple structural determinants. While our analysis focuses on Twitter, we view our analysis of social-topical affiliations as broadly applicable to a host of diverse affiliations, including the movies people watch, the brands people like, or the locations people frequent.

The chapters in this thesis are based on papers that have been published in recent years [59, 113, 116, 114, 112, 115]

## CHAPTER 2

### LITERATURE REVIEW

There is growing body of research about online social and information networks. Typically, these networks are abstractly represented by graphs. Graphs are used to systematically study various properties of the network. In the simplest case, we assume that people or pieces of information are nodes and undirected links between nodes signal that there is a relationship between them. This simple abstraction allows us to begin to understand the structure of actual social and information networks, how they form, and the way in which information travels through their links. In this section, we will review some of the theoretical and empirical findings related to link formation and information diffusion.

## 2.1 Generative Models for Graph Formation

We will begin by discussing some basic graph generative models for social and information networks. These models are usually motivated by sociological theories and by empirical observations from real data.

### 2.1.1 Small-World Networks

In the 1960s, Milgram published the results of a series of experiments he conducted which revealed an interesting property of the structure of the connections between people around the world [123, 99]. The experiments consisted of giving a letter to randomly selected subjects in various cities across the United

States and asking them to forward the letter to anyone they personally knew on a first name basis. The person they forwarded the letter to would receive the same set of instructions. Subjects were told that the goal was for the letter to eventually reach a person who lived in Boston and they were given the basic contact information of this person. In one of his experiments, Milgram gave the letter to 296 randomly selected people in Nebraska. Out of the 296 letter only 64 reached the target person in Boston. The surprising part is that the 64 letters that reached the target, only took an average of 6 hops to go from the source to the target. This finding gave rise to the popular phrase "six-degrees of separation," which suggests that the average path between two people in a social network has length 6.

The idea that we are all connected by relatively short paths has motivated research on generative models of graph formation that have this property. One of the most well-known models of this kind is the Watt-Strogatz model for Small-World Networks [128]. The Watts-Strogatz model starts with a set of nodes evenly spread on a circle with each node connected to its  $k$  closest neighbors. Every original edge from each node is then rewired to a random node with probability  $p$ . The parameter  $p$  allows for the amount of randomness in the network to be tuned between all random edges ( $p = 1$ ) and no random edges ( $p = 0$ ). Watts and Strogatz pose that when  $p$  is between 0 and 1, the model gives rise to networks with two important ingredients that are present in real social networks. One ingredient is that there exist small paths among randomly selected nodes as Milgram found in his experiments. These short paths exist because when  $p > 0$  some of the original short-range edges are rewired, and they become long-range edges, which connect nodes that are far away from each other in the circle. The second ingredient is high *clustering coefficient*, which means

that there is a tendency for nodes that have a neighbor in common to also be connected. This happens because, in the original graph, as long as  $k > 0$ , some of the neighbors of each node will also be neighbors. Since  $p < 1$ , this property is conserved. Watt-Strogatz empirically show that real world networks, such as a network of actors appearing in the same film, the power grid in the western US, and the neural network of the nematode worm *C. elegans*, have this two ingredients as well.

Milgram's experiments did not only show evidence that there are short paths between randomly selected people. They also showed that people can find such paths using only the local information they have about the network. The Watts-Strogatz model gives rise to networks that have these short paths, but it does not explain why people on the network are able to find them using only local information. Through a generalization of the Watt-Strogatz model, where a parameter  $r$  controls how far the long-range edges can go when they are rewired, Kleinberg showed that decentralized algorithms can effectively route messages through short paths in these networks only for a particular parameter  $r^*$  [69]. The intuition is that when long-range edges cannot extend far enough then it takes many small hops to get the message to the target. However, when long-range edges can extend arbitrarily far, as in the original Watts-Strogatz model, even though short paths exist, they cannot be efficiently discovered with just local information. That is because as the message gets closer to the target, no long-range edges are likely to link the current holder of the message to a node near the target. Hence, when long-range edges can extend arbitrarily far, they cannot be used to route the message.

### 2.1.2 Preferential Attachment

Another observation that has been made about real world networks, including social ones, is that their degree distribution approximates a power law. That is, the number of nodes with degree  $k$  is proportional to  $\frac{1}{k^\alpha}$  [10, 14]. This distribution has been found to describe quantities that measure popularity in other domains such as the number of citations of scientific papers and the number of calls received by telephone numbers [102]. The most basic generative model that roughly gives rise to this degree distribution is the preferential attachment model, which was originally proposed by Price in 1976 [105], and later applied to model the growth of the World Wide Web [10]. When the model is applied to the formation of networks, the network starts out as node with an edge from and to itself, then each additional node sequentially joins the network and creates a link to a random node such that the probability of linking to a node  $v$  is proportional to the current degree of  $v$ . While this model produces networks with a degree distribution similar to the ones found in real networks, it fails to produce networks with as much clustering as the ones produced by the Watts-Strogatz model. On the other hand, the degree distributions of the networks produced by the Watts-Strogatz model are not power laws.

Extensions of the preferential attachment model have been proposed and we discuss some of them in chapter 4. In particular, a preferential attachment model that generates graphs with more clustering than the original one is the “copying model” [74, 102, 124]. In one version of this model, when a node joins the network it links to a node  $v$  uniformly at random, and then chooses some of the neighbors of  $v$  to link to as well. The idea of the model is that the node selects a random node and then *copies* some of its connections. Notice that the

new node is more likely to attach to a node with high in-degree than to one with low in-degree. Also, since most of the links formed close triangles, the clustering of the resulting graph is higher than in the original preferential attachment model. Other models related to preferential attachment are preferential attachment with fitness, where each node  $v$  has a “fitness parameter”  $f_v$ , and when a new node joins the network, it links to  $v$  with probability proportional to  $f_v * d_v$ , where  $d_v$  is the degree of  $v$  [11].

## 2.2 Beyond Static Simple Graphs

Real social networks are very complex structures. As we have seen, representing them with simple graphs allows us to study important structural properties such as degree distributions, shortest paths between nodes, and clustering. However, other key properties are missed by this representation. In real social networks, relationships are not symmetric, they are not all of the same kind or intensity, and they are not constant. In this section we review some generalizations of simple graphs that allow us to capture some of these features of real social networks that are missed by static simple graphs.

### 2.2.1 Weighted Graphs

Social relationships vary in their intensity. If we measured the amount of yearly communication a person has with each of the people she knows, we would naturally find that there are some people she communicates with numerous times such as close family members, close friends, and co-workers, and there are oth-



ers she communicates with very few times such as old high school friends or acquaintances. When we represent social networks with unweighted graphs we assume that all links are equal. This assumption could hide aspects of the network that impact the dynamics of what we are aiming to study. For example, one reason why we study social networks is that they naturally play a role in how information is diffused from person to person. If we assume all ties have the same strength we would be missing important differences in the role that ties of different weights play in the diffusion of information. As we would expect, strong ties play an important role in information diffusion because of their high levels of communication, what is somewhat more surprising is that weak ties also play an important role.

In the 1960s, Granovetter studied the importance of weak ties in the diffusion of information and influence. In one of his studies, he interviewed people who had recently found a job. An interesting and perhaps counterintuitive finding was that many of the subjects indicated that they learned about their job through acquaintances and not from close friends [45]. Granovetter theorized that the reason for this is that a person's close friends are likely to be in the same social circle as the person and hence they are all exposed to the same information. On the other hand, a person's acquaintances serve as bridges between social circles and carry new information from one circle to the other [44]. Other research has shown that weak ties are important in the diffusion of good ideas [19].

While it is difficult to define weak and strong ties precisely, there has been work on the measurement of tie strength [94]. Data from social networks often provide enough information to come up with reasonable proxies for measuring the strength of relationships. For example, weak ties have been defined through

the number of triangles they close [118], whether they are reciprocal or not [92], and by the amount of communication that is exchanged through them [37]. Using various definitions of weak and strong ties, it has been found that weak ties have an impact in the overall structure of a social network. By analyzing data from Facebook, researchers have found that ties vary in strength when they are measured by whether communication between nodes is reciprocal and maintained [92]. They also found that when looked at separately, weak and strong ties form networks of different tie density. However, in other domains, it has been found that when the strength of a tie is measured by the number of triangles it closes, removing weak ties does not cause a large increase in the diameter of the network and it does not reduce the size of its largest connected component by a significant amount. We will discuss related empirical findings on Twitter in chapter 3.

With the vast amount of information that is presented to users of social media sites everyday, it becomes important to be able to automatically detect tie strength. Designer of these sites have the challenge to avoid overloading users with information about far away acquaintances they are not interested in and, at the same time, allowing information to flow through different parts of the social network. As we know, weak ties play a central role in achieving this goal. In this direction, there has been work on the prediction of tie strength [35, 25, 108].

### **2.2.2 Signed Graphs**

Even if we allow for edges to carry a weight when we think of social networks as graphs, we are still assuming that all edges have the same general meaning,

just with different strengths. However, social networks have different kinds of edges. For example, not all relationships in a social network are friendly. If one wanted to study social networks where people can be friends or enemies, it would not be enough to allow edges to have a positive weight, one would need to allow for signed edges.

One theory related to signed social networks says which triads are likely to exist in a signed social network [53, 20, 127]. Balance theory poses that triads with zero or two negative edges are *balanced* and more likely to exist than other triads, while triads with one or three negative edges are *unbalanced* and likely to only exist for a short time. The intuition for why balanced triads are more likely to exist is the idea that a friend of my friend is my friend (no negative edges), and that the friend of my enemy is my enemy (two negative edges). On the other hand, when there is only negative edge, one node  $v$  in the triad has two friends who are each other's enemy, which creates stress on  $v$ . In this case, the theory says that either the negative edge will become positive, or  $v$  will become an enemy of one of her previous friends. In either case, the triad will become balanced. In the case where there are three negative edges, the theory says that the two nodes with the least conflict are likely to become friends to join forces against the third node, making the triad balanced.

Large scale validation of balance theories have been difficult to perform due to lack of large signed network data sets. However, with the richness of data from social media sites, progress has been made in this direction. In particular, a study based on data from three different sites, where edges are naturally signed, found that triads with only one negative edge are indeed significantly less frequent than expected by random chance, as balance theory would suggest

[84]. However, they also found that triads with all three negative edges are *more* frequent than expected by random chance, which contradicts balance theory. Their findings are consistent with a slightly different version of balance theory which labels triads with exactly one negative edge as unbalance and all others as *weakly* balanced [29]. Another study shows that balance theories, along with related theories applied to status instead of friendship, can be useful to develop techniques to predict positive and negative links [83].

Other large scale results of signed networks include a study which found that positive and negative relationships have a different effects in what and how users will vote in an online social network where users can vote on the opinions of others and can declare allies and nemeses [16]. Another study proposed a generalization for various measures of signed networks that are typically applied to networks of only positive edges such as signed clustering coefficient and negative rank [75].

### 2.2.3 Evolving Graphs

Much of the empirical research on online social networks has focused on taking a snapshot of the network and studying its current state without considering how it looked in the past or how it will look in the future. In this way, much progress has been made in understanding the basic structure of the network, but less can be said about the fundamental mechanisms that control how these networks are formed. As more detailed networked data become available for research, people have started to pay more attention to the evolution of social networks. The study of the evolution of social networks is a fundamental as-

pect of link prediction [87, 104, 117], which plays a major role in the creation of successful link recommendation algorithms for social media sites [24, 42, 17]. Recent work has studied how different sociological theories play a role in the formation of real social networks. Two basic theories related to the formation of social links are *homophily* and *triadic closure*.

One of the most basic principles of link formation is *homophily*, the idea that people who have social connections tend to be similar to each other [97, 78]. Empirical studies have found homophily to be present in snapshots of social networks, evidenced by high demographic similarity among connected pairs in a friendship network of middle school students [100]. Furthermore, studies have found that homophily plays an important role in the formation of a network friendships among students of a university [72]. There are many mechanisms that can give rise to homophily in social networks and it's not easy to tell which particular mechanism is actually responsible for it. For example, it is possible that connected people are not similar to each other when they first meet, but become similar because of the influence they exert on each other. Another possibility is that people are more likely to meet others who are similar to them, and hence, connected people are similar to each other from the moment they meet. Methodologies for determining which of these two mechanism has a stronger impact in the formation of social networks have been proposed [27]. However, results have varied depending on the particular social network studied, which suggests that both mechanisms are present to some extent [64, 27].

The idea of homophily gives rise to more complex mechanisms of link formation. For example, if similar people are likely to form relationships, and connected people tend to be similar, then by transitivity of similarity, if two people

have a friend in common, they are likely to be similar. Furthermore, since similar people tend to connect, then the two people with a friend in common should tend to connect. This is what the principle of *triadic closure* says, that people with friends in common tend to become friends [107]. Triadic closure has been empirically shown to be a present mechanism in the evolution of social networks [71]. In chapter 4 we will define a generalization of triadic closure for directed networks and measure its effect in the evolution of social-information networks.

## 2.3 Information Diffusion

Spread of information is an increasingly popular subject of study and social networks play a central role in this line of research. Information diffusion in social networks refers to the idea that people have a tendency to perform an action, adopt a belief, or simply become aware of something after some of the people they know have done so. From the point of view of Sociology, the diffusion of information has been studied both theoretically and empirically [110, 119]. Furthermore, diffusion processes happening online such as news propagation, viral marketing of many products, and spread of political ideologies have also been studied [79, 2, 76, 86, 48]. In this section we review some basic findings related to information diffusion.

### 2.3.1 Basic Models of Information Diffusion

Two basic models for information diffusion are the *linear threshold model* and the *independent cascade model*.

**Independent Cascade Model** A simple version of the independent cascade model [40] assumes the diffusion of a piece of information or behavior in an undirected graph happens in the following way. At every time step  $t$ , each node is either *infected* or not. Here, being infected corresponds to having become aware the piece of information or having adopted the behavior. At every time step  $t$ , each node that became infected in step  $t - 1$  will get one opportunity to infect its neighbors. A node  $v$ , attempting to infect its neighbor  $u$ , will be successful with probability  $p(v, u)$ , which is fixed for all pairs of neighbors at the beginning of the process. The process ends when no new nodes become infected. Generalizations of this model have been proposed where the infection probabilities change throughout the process in order to make the model more realistic. For example, the  $p(v, u)$  may depend on how many times and which nodes have attempted to infect  $u$  [65].

**Linear Threshold Model** The linear threshold model [46] considers the effect that the whole set of infected nodes has on influencing each node  $v$ . In a version of the model, each node  $v$  has a function  $T_v$  that takes as input any set of nodes  $I$  and outputs a number in  $[0, 1]$ . Node  $v$  also has a threshold  $z_v$ . At each time step, each node  $v$  will become infected if  $T_v(I) > z_v$ , where  $I$  is the set of currently infected nodes.

The linear threshold and independent cascade models are intrinsically different because one considers the effect that each individual node has in influencing or infecting a node  $v$ , and the other one considers the effect of the infected nodes as a whole. However, a generalized version of the independent cascade model has been shown to be in some sense equivalent to the linear threshold model. That is, for any choice of functions  $T_v$  in the linear threshold model, there exist a

set of parameters for the generalized independent cascade model such that for any set of nodes  $N$  and any time step  $t$ , the probability that the set of infected nodes at time  $t$  is the set  $N$ , is the same for both models [65].

### 2.3.2 Online Diffusion Processes

Data from online diffusion processes have become available to researchers in recent years. Being able to access data at large scale has allowed researchers to better understand the dynamics of information diffusion, to measure influence, and to validate theoretical models on real data.

Work on the visualization [3] and measurement of real cascades of information spread can begin to provide a picture of the general dynamics of the process. Data from a chain letter, which spread widely over the internet, produced a cascade which did not fan out significantly, instead it created long and narrow chain configurations [89]. On the other hand, it has been found that cascades created by messages being forwarded on Twitter tend to be very short [9]. This suggests that the dynamics of diffusion are dependent on the domain in which the process is taking place and also on the topic of the information. It has also been shown that the propensity that people have to join LiveJournal communities and research communities does not only depend on how many of their neighbors have joined the community, but also on how the neighbors are connected among each other [8]. Similarly, it has been found that the number of friends a person has on Facebook does not provide a significant signal for predicting the size of an information cascade initiated by the person [120]. These results indicate that the independent cascade model, where each node has a cer-



tain probability of influencing another without regard of who else is infected, may not be appropriate in all domains. There has been work on designing models that produced similar diffusion cascades of product recommendations [79] and blog posts [86].

### 2.3.3 Influence

Influence is at the heart of information diffusion in social networks. The basic assumption of information diffusion is that people who have received the information or adopted a certain behavior, referred to as infected nodes, exert some amount of influence on others so that those connected to infected nodes have an increased probability of becoming infected. Methodologies for measuring levels on influence on a social network based on diffusion processes have been proposed [26]. The authors posit that the most basic measure of influence is the function  $f(k)$ , which gives the average probability that a node  $v$  will become infected given that  $k$  neighbors are infected. They propose two ways of constructing the function  $f(k)$ , one based on a series of snapshots of the state of the network including who is currently infected, and another way based on complete data with timestamps of when each node became infected. In chapter 6, we use this methodology to study the differences of information diffusion dynamics by topic.

Some work has focused on measuring influence probabilities. In other words, measuring the probability  $p(v, u)$  that a node  $v$  can influence its neighbor  $u$  to adopt certain behavior [43]. However, as discussed earlier, the probability that a node can be influenced may not only depend on just the number

of neighbors that have adopted the behavior, but rather on the whole set of infected nodes. Therefore, other work has focused on identifying influential nodes in the network, those who make the most impact in the probability that other nodes will become infected when they are in the set of infected nodes. In this direction, models for quantifying the overall influence that bloggers have in spreading information on the Web have been developed [4]. Also, there has been work on algorithms for identifying influential topic-specific bloggers [129]. In chapter 7, we propose an algorithm to rank users of a social media site by the influence they have in, not only propagating information, but also getting other users to become active participants of the site.

It is important to keep in mind that all the mechanisms of network formation and information diffusion we have discussed affect the network at the same time, and it can be hard to distinguish which mechanism is at work when we make observations from data. For example, one may observe that when a person buys a product, some of its friends do so as well. It is hard to determine if this observation is due to the fact that friends tend to be similar and like the same products, or if it's due to the influence that friends exert on each other. In other words, it's not clear if the observation is due to homophily or influence. Methodologies for distinguishing the effects of homophily and influence have been proposed [7].

## **Part I**

# **Network Structure Evolution**

## CHAPTER 3

### SOCIAL-INFORMATION NETWORKS

Social networks, a very old and pervasive mechanism for mediating distal interactions among people, have become prevalent in the age of the Web. With interfaces that allow people to follow the lives of friends, acquaintances and families, the number of people on social networks has grown exponentially since the turn of this century. Facebook, LinkedIn and MySpace, to give a few examples, contain millions of members who use these networks for keeping track of each other, find experts and engage in commercial transactions when needed [70]. Furthermore, commercial enterprises try to exploit them for marketing purposes, as they provide a ready made medium for propagating recommendations through people with similar interests [79].

While the standard definition of a social network embodies the notion of all the people with whom one shares a social relationship, in reality people interact with very few of those “listed” as part of their network. One important reason behind this fact is that attention is the scarce resource in the age of the web. Users faced with many daily tasks and large number of social links default to interacting with those few that matter and that reciprocate their attention. For example, a recent study of Facebook showed that users only poke and message a small number of people while they have a large number of declared friends [41]. And a casual search through recent calls made through any mobile phone usually reveals that a small percentage of the contacts stored in the phone are frequently contacted by the user.

These initial observations suggest a systematic investigation into the nature of the social networks where there is large variation in the amount of interaction

among connected nodes. An interesting question is whether the network that is made out of the pattern of interactions that people have with their friends or acquaintances and the one that is constructed from a list of all the contacts they may decide to declare are two different networks in nature, or just two social networks with edges that have different strength.

In order to construct and compare these networks, we collected and analyzed a large data set from the Twitter social network. Twitter.com is an online social network used by millions of people around the world to stay connected to their friends, family members and coworkers through their computers and mobile phones. The interface allows users to post short messages (up to 140 characters) that can be read by any other Twitter user. Users declare the people they are interested in following, in which case they get notified when that person has posted a new message. A user who is being followed by another user does not necessarily have to reciprocate by following them back, which makes the links of the Twitter social network directed.

### 3.1 Results

For each user of Twitter in our data set we obtained the number of *followers* and *followees* (people followed by a user) the user has declared, along with the content and datestamp of all his posts.<sup>1</sup> Our data set consisted of a total of 309,740 users, who on average posted 255 posts, had 85 followers, and followed 80 other users. Among the 309,740 users only 211,024 posted at least twice. We

---

<sup>1</sup>Twitter only displays up to 3201 updates per user so we only have the complete set of updates for users who have posted 3200 or less updates. A very small set of users showed 3201 updates so we have the complete set for about 99.6% of all the users.

call them the *active users*. We also define the *active time* of an active user by the time that has elapsed between his first and last post. On average, active users were active for 206 days.

Twitter users are able to post direct and indirect updates. Direct posts are used when a user aims her update to a specific person, whereas indirect updates are used when the update is meant for anyone that cares to read it. Even though direct updates are used to communicate directly with a specific person, they are public and anyone can see them. Often times two or more users will have conversations by posting updates directed to each other. Around 25.4% of all posts are directed, which shows that this feature is widely used among Twitter users.

We are interested in finding out how many people each user communicates directly with through Twitter. We define a user's *friend* as a person whom the user has directed at least two posts to. Using this definition we were able to find out how many friends each user has and compare this number with the number of followers and followees they declared.

Based on our previous finding about the role of attention in eliciting productivity within a social network [58], we conjecture that the users who receive attention from many people will post more often than users who receive little attention. Therefore we expect that users with more followers and friends will be more active at posting than those with a small number of followers and friends. Figures 3.1 and 3.2 show that indeed the total number of posts increases with both the number of followers and friends. However, as figure 3.1 shows, the number of total posts eventually saturates as a function of the number of followers. This implies that users with a large number of followers are not necessarily

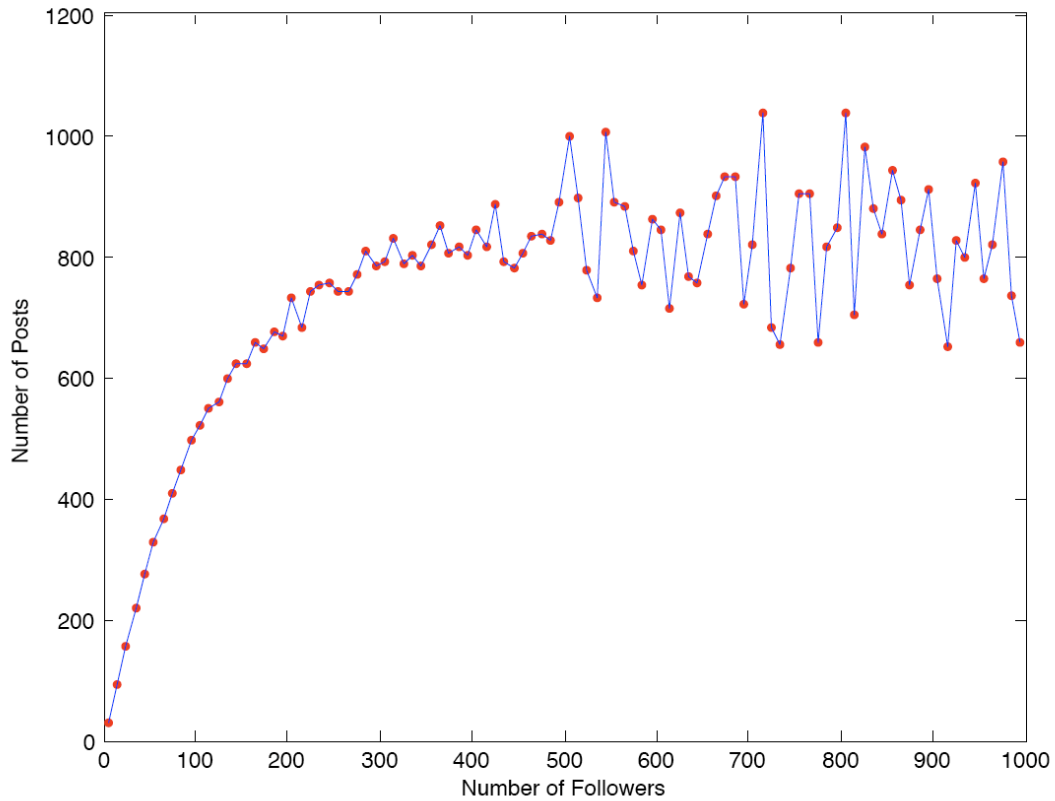


Figure 3.1: Number of posts as a function of the number of followers. The number of posts initially increases as the number of followers increases but it eventually saturates.

those with very large number of total posts. On the other hand, the number of total posts does not saturate as a function of number of friends, as seen on figure 3.2. Rather, the number of updates increases until it reaches a maximum point of 3201. This suggests that in order to predict how active a Twitter user is, the number of friends is a more accurate signal than the number of his followers.

Having shown that the number of friends is the actual driver of Twitter user's activity, we compared it with the number of followees the users declare. We define  $\delta$  as the number of friends a user has, divided by the number of fol-

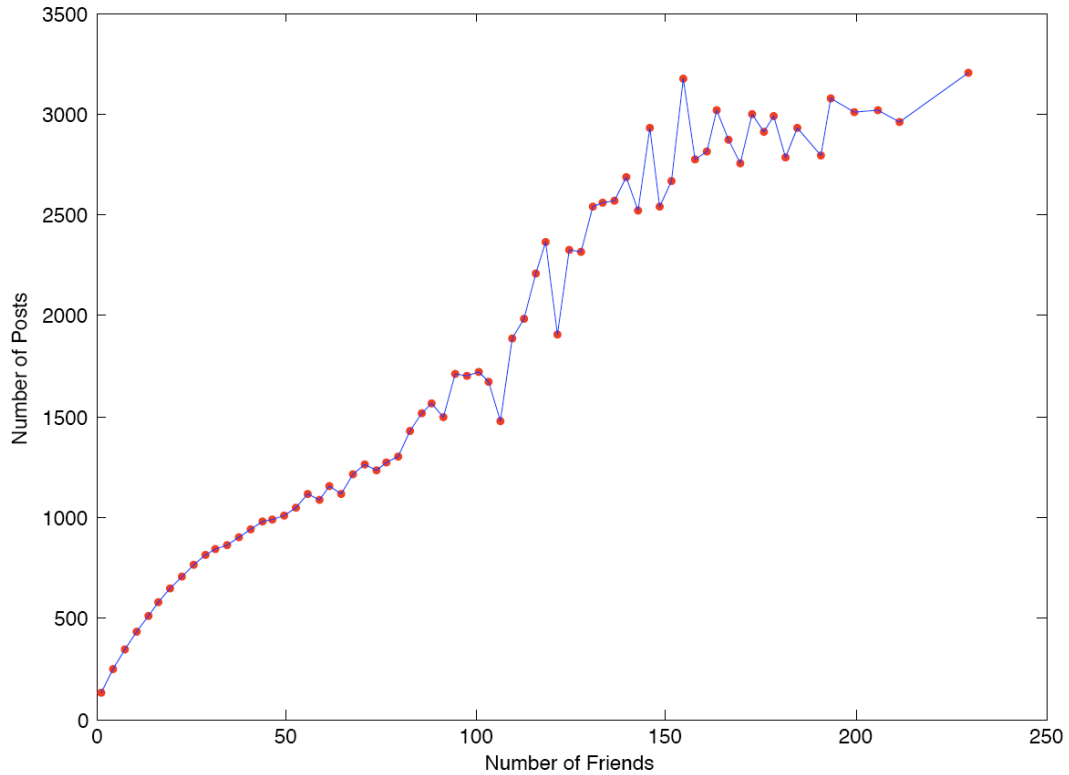


Figure 3.2: Number of posts as a function of the number of friends. The number of posts increases as the number of friends increases, reaching 3200 without saturating.

lowees she declared. Since 98.8% of the users have fewer friends than followees, almost all the  $\delta$  values are less than 1. Figure 3.3 shows a histogram of the  $\delta$  values. As we can see most users have a  $\delta$  value less than .1, with the number of users with a  $\delta$  close to 1 extremely small. The average of the  $\delta$  values is 0.13 and the median is 0.04. This indicates that the number of friends users have is very small compared to the number of people they actually follow. Thus, even though users declare that they follow many people using Twitter, they only keep in touch with a small number of them. Hence, while the social network created by the declared followers and followees appears to be very dense, in reality the



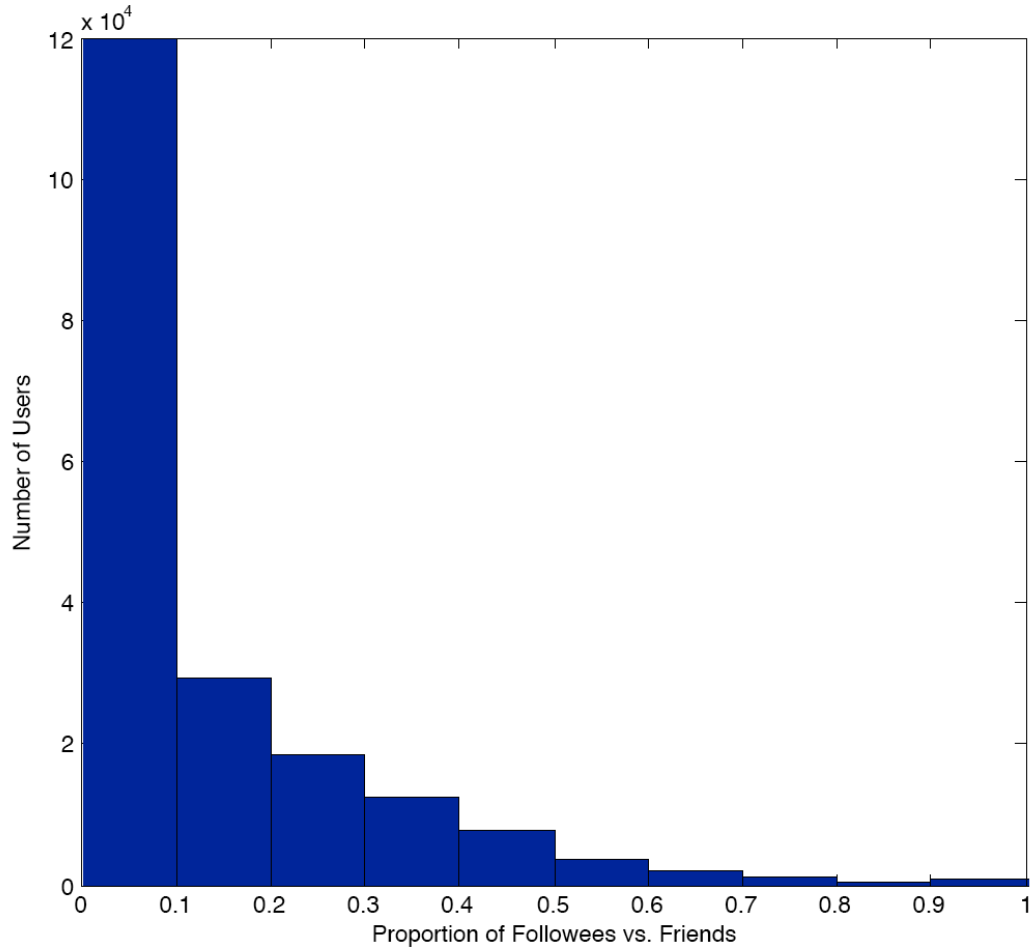


Figure 3.3: Histogram of contributor's number of friends divided by the number of followees. Most users have a very small number of friends compared to the number of followees they declared.

network of friends suggests that the social network is sparse.

Another interesting aspect is to consider how the number of friends and the  $\delta$  values change as the number of followees increases. Figures 3.4 and 3.5 show that even though the number of friends initially increases as the number of followees increases, after a while the number of friends starts to saturate and stays nearly constant. This trend can be explained by the fact that the cost of

declaring a new followee is very low compared to the cost of maintaining a friends (i.e. exchanging directed messages with other users). Hence, the number of people a user actually communicates with eventually stops increasing while the number of followees can continue to grow indefinitely.

There are two other way of interpreting figure 3.4. If we think of the weak ties of a user as those who she follows but does not communicate with, and strong ties as those who she follows and communicates with, then figure 3.4 says that, while social media sites allows people to have an arbitrary number of weak ties, the maximum number of strong ties is bounded. This validates theories that posit that a person is only able to maintain a limited number of social relationships [31]. The other way to explain figure 3.4 is to think of follow relationships and “friend” relationships as different in nature. Twitter users follow not only their friends but also celebrities, politicians, news generators, and other organizations. It is possible that those accounts a users follows but does not communicate with tend to belong to organizations as opposed to people. In this case, we could think of Twitter as a social-information network where ties sometimes signal social relationships, but other times they signal that a node is simply interested in another node but they do not actually know each other.

## 3.2 Discussion

Even when using a very weak definition of “friend” (i.e. anyone who a user has directed a post to at least twice) we find that Twitter users have a very small number of friends compared to the number of followers and followees they declare. This implies the existence of two different networks: a very dense one

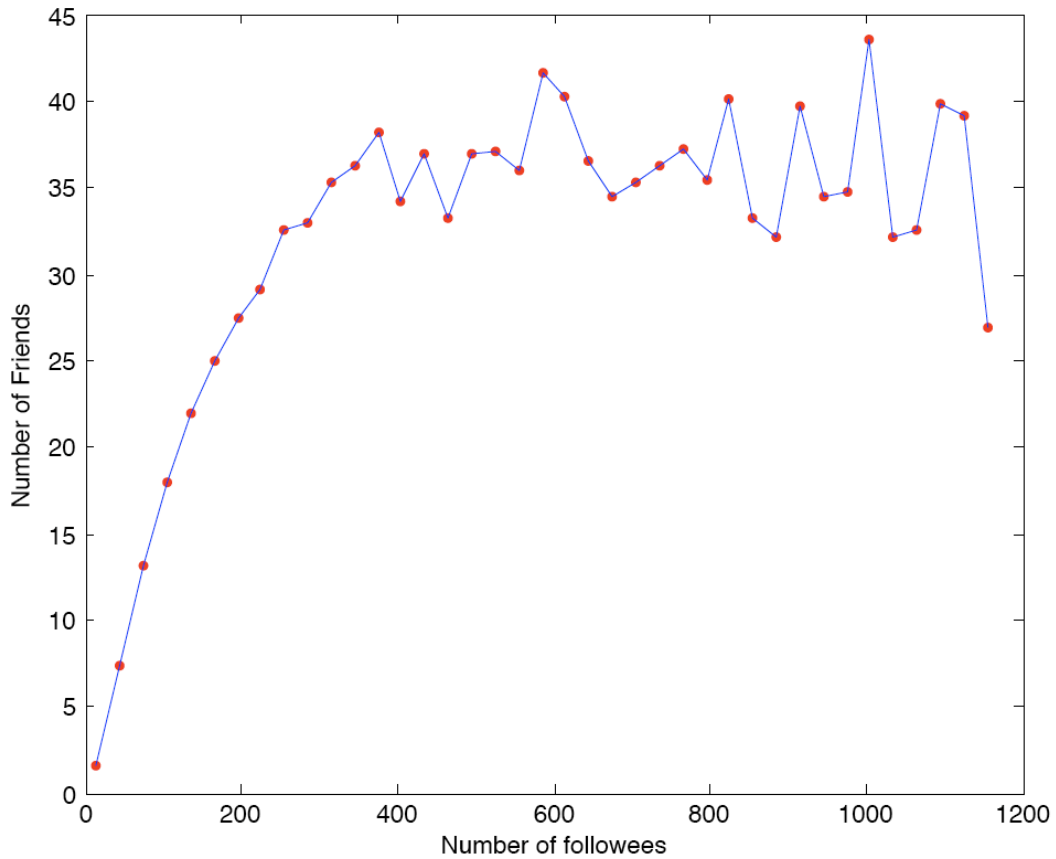


Figure 3.4: Number of friends as a function of the number of followers. The total number of friends saturates while the number of followers keeps growing due to the minimal effort required to add a followee.

made up of followers and followees, and a sparser and simpler network of actual friends. The latter proves to be a more influential network in driving Twitter usage since users with many actual friends tend to post more updates than users with few actual friends. On the other hand, users with many followers or followees post updates more infrequently than those with few followers or followees.

We find that users are able to accumulate many declared connections on

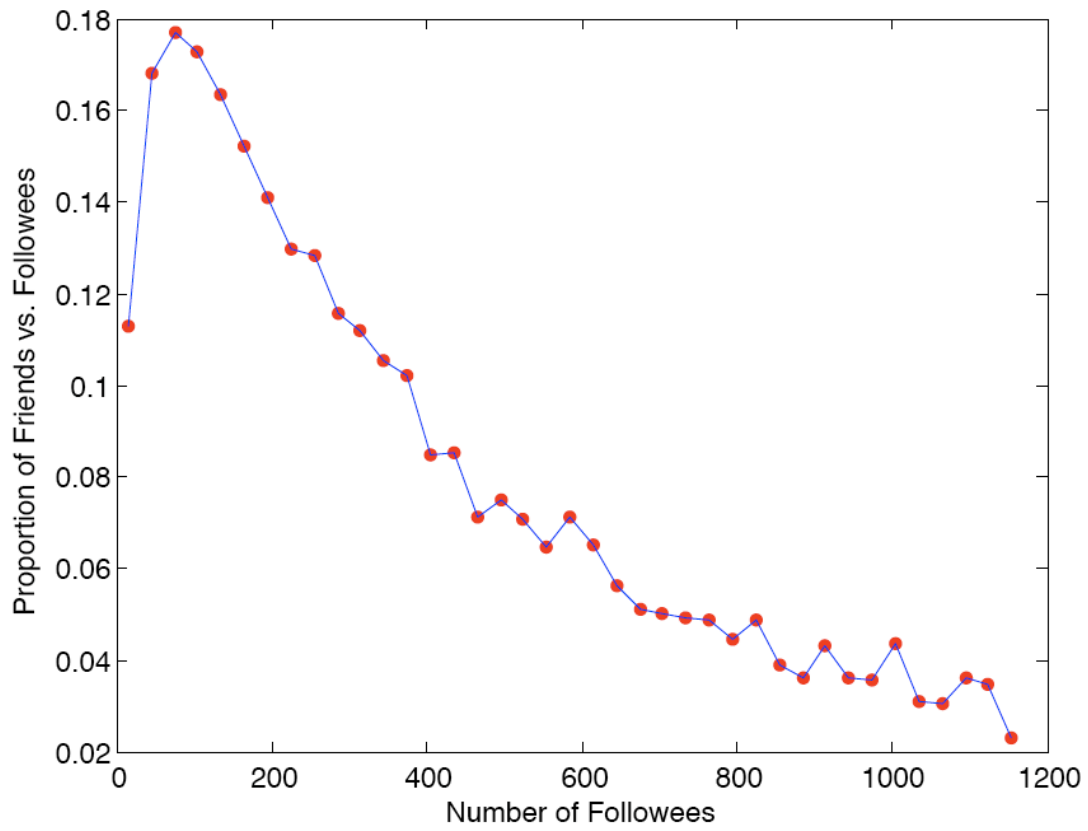
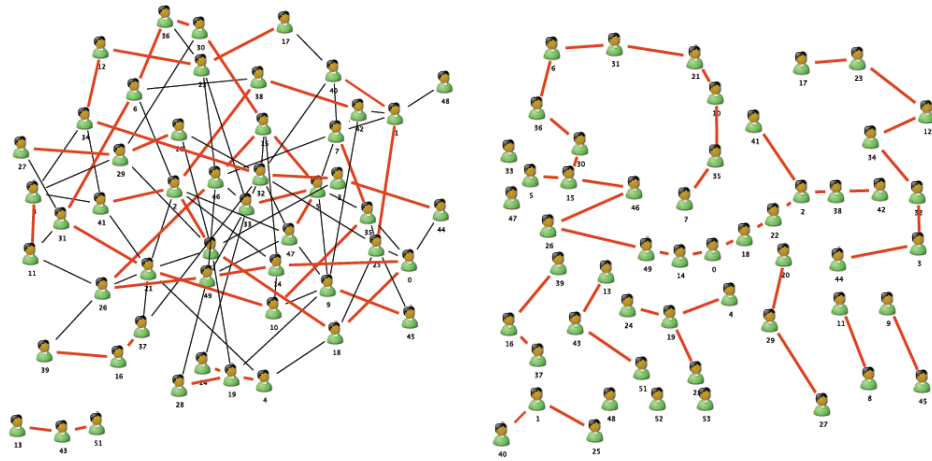


Figure 3.5: Proportion of friends vs. followers as a function of followers. It initially increases but rapidly approaches zero as the number of followers increases.

Twitter. However, they only maintain communication with a small number of them. We propose two different explanations for this finding. One is that people can have many acquaintances but the number of close friends they can manage to maintain is bounded. And the other is that Twitter is not just a social network but a social-information network, where people connect and interact with their friends but also connect to organizations they are interested in.

Many people, including scholars, advertisers, and political activists, see on-line social networks as an opportunity to study the propagation of ideas, the formation of social bonds and viral marketing, among others. This view should



(a) All links are declared followees and the red links are actual friends.  
 (b) After removing the black links and reorganizing the network look simpler than before. This is the hidden network that matters the most.

be tempered by our findings that a link between any two users does not necessarily imply an interaction between them. As we showed in the case of Twitter, most of the links declared within Twitter were meaningless from an interaction point of view. Thus the need to find the hidden social network; the one that matters when trying to rely on word of mouth to spread an idea, a belief, or a trend.

## CHAPTER 4

### THE DIRECTED CLOSURE PROCESS IN SOCIAL-INFORMATION NETWORKS

Information networks, which connect Web pages or other units of information, and social networks, which connect people, are related notions, but they exhibit fundamental differences. Two of the principal differences are based on directionality and heterogeneity. First, information networks are generally directed structures, with links created by one author to point to another; social networks, on the other hand, tend to be represented in most basic settings as undirected structures, expressing relationships that are approximately mutual. Second, information networks tend to contain a few nodes with extremely large numbers of incoming edges — documents or pages that are “famous” and hence widely referenced — while social networks exhibit disparities in connectivity only to a smaller extent, since even the most gregarious people have some practical limit on the number of genuine social ties they can form.

The link structure of the Web, and of well-defined subsets of the Web such as the blogosphere and Wikipedia, are clear examples of information networks; social-networking sites such as Facebook have provided us with very large representations of social networks that are derived from social structure in the off-line world. An interesting recent development has been the growth of social media sites that increasingly interpolate between the properties of information networks and social networks. The micro-blogging site Twitter is a compelling example of such an interpolation.

As we discussed in chapter 3, the structure of the Twitter network reflects properties both of a social network, since it exposes underlying friendship re-

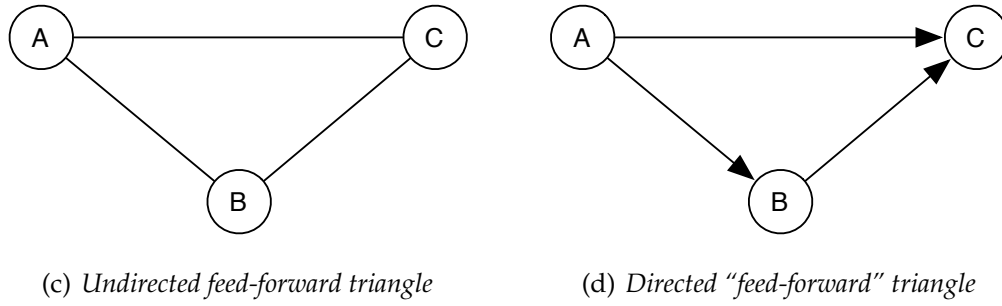


Figure 4.1: (a) Triadic closure in an undirected graph produces a triangle when an edge connects two nodes who already have a common neighbor. (b) Analogously, in a directed information network, directed closure occurs when a node  $A$  links to a node  $C$  to which it already has a two-step path (through a node  $B$ ). This creates a directed triangle (a “feed-forward” structure on three nodes).

lations among people, and also of an information network, since it is directed and also contains huge concentrations of links to specific “celebrities” and automated generators of news content that reflect fundamentally informational relations.

**Link Formation in Information Networks.** In a social network, triadic closure is one of the fundamental processes of link formation: there is an increased chance that a friendship will form between two people if they already have a friend in common [107, 44]. (For example, we could imagine the  $A$ - $C$  friendship in Figure 4.1(c) as forming after the existence of the  $A$ - $B$  and  $B$ - $C$  edges, and accelerated by the existence of these two edges.) Recent empirical analysis has quantified this effect on large social network datasets [71]. Is there an analogous process in information networks?

A natural hypothesis for such a process is the following: if a node  $A$  in an information network links to  $B$ , and  $B$  links to  $C$ , then one should arguably ex-

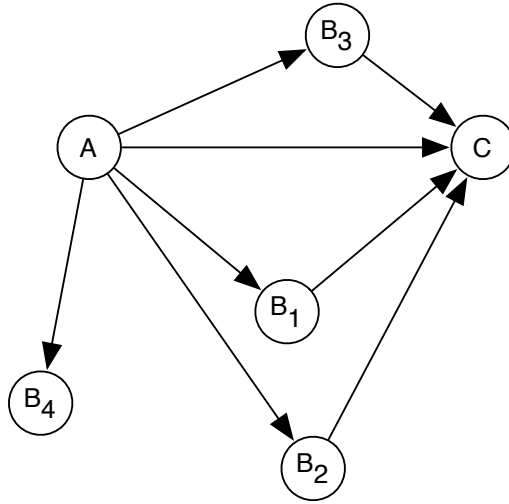


Figure 4.2: In this example, the edge from  $A$  to  $C$  exhibits closure if there is already a two-step path from  $A$  to  $C$  (i.e., through  $B_1, B_2, B_3$ ) when the  $A$ - $C$  edge arrives.

pect an increased likelihood that  $A$  will link to  $C$  — since the author of  $A$  has an increased ability to become aware of  $C$  via the two-step path through  $B$ . (See Figure 4.1(d).) We will refer to this as the *directed closure process*. In addition to its intuitive appeal, this process contains an implicit hypothesis about how links are formed in information networks — through the “copying” of a link from something you already point to — and such copying mechanisms form a crucial part of the motivation for the fundamental notion of preferential attachment [5, 74, 102]. Despite the importance of the notion, however, there has been remarkably little empirical analysis of the extent to which this type of directed closure is truly at work in real information networks, and of the effects it may have on network structure.

**The Directed Closure Process.** In this chapter, we analyze the directed closure process using data from Twitter: we provide some of the first evidence on large



information networks that directed closure is taking place at a rate significantly above what would be expected by chance; we identify a surprising level of heterogeneity in how strongly it operates across different parts of the network; and we analyze models that capture these effects.

An important difference between triadic closure in social networks and directed closure in information networks is the following observation, which in a sense serves as the starting point for our analysis: while the extent of triadic closure can be assessed from a single snapshot of an undirected graph, the evaluation of directed closure inherently requires some form of temporal sequence information. Indeed, when we see an undirected triangle such as the one in Figure 4.1(c), we know that whichever edge formed last will complete a two-step path consisting of the earlier two edges, and hence will satisfy the definition of triadic closure. On the other hand, the structure in Figure 4.1(d) satisfies the definition of directed closure only if the  $A$ - $C$  edge formed after the other two.

This means that the amount of directed closure in a directed graph depends not just on the graph's structure, but also on the order in which edges arrive. Because of this we are able to develop a natural *randomization test* to evaluate whether directed closure is taking place in a given network at a rate above chance. Specifically, we say that an edge in a directed graph *exhibits closure* if, at the time it forms, it completes a directed two-step path between its endpoints. For example, in Figure 4.2, the  $A$ - $C$  edge exhibits closure if and only if it arrives after the pair of edges in one of the three possible two-step  $A$ - $C$  paths through  $B_1$ ,  $B_2$ , or  $B_3$ . For a given network, we can thus ask: how many edges exhibit closure, and how many would have exhibited closure (in expectation) if the edges had arrived in a random order? The point is that in any arrival order of the

edges, some number of the edges will close directed triangles; but if directed closure is a significant effect, then we may expect to see a larger number of such triangle-closings compared to what we'd see under a random arrival order.

To investigate this empirically, we choose a random sample of *micro-celebrities* on Twitter, which we define to be users with between 10,000 and 50,000 followers. (We will abbreviate the term as  $\mu$ -celebrity.) For each such  $\mu$ -celebrity  $C$ , we determine the number of edges to  $C$  that exhibit closure, and compare it to the expected number of edges to  $C$  that would exhibit closure in a random ordering — we will refer to this latter number as the *random-ordering baseline*. Given that we are studying the followers of users with high numbers of in-links, one would conjecture that there are two competing forces at work. In one direction is the intuitively natural tendency of directed closure to create short-cuts in the presence of two-step paths. In the other direction, however, is the plausible tendency for people to link first to celebrities, before they link to more obscure users; that is, it is not clear that closure processes are necessary in order for people to discover and link to very prominent users. This latter effect would tend to cause triangles as in Figure 4.1(d) to appear with the  $A$ - $C$  and  $B$ - $C$  edges first, reducing the extent of directed closure in the real data.

We find in the Twitter data that the number of edges to a  $\mu$ -celebrity that exhibit closure is higher than the random-ordering baseline, indicating that even in linking to celebrities, there is an above-chance tendency to do this by closing an existing two-step path. This finding suggests a range of further interesting questions — specifically, whether the high rate of directed closure is due to overt copying of follower lists (as in the intuitive basis for the definition), or due to more subtle, implicit mechanisms that produce copying behavior at a macro-

scopic level. To address this question, as we discuss below, we consider the extent to which directed closure can arise even in models that do not explicitly build in copying as a mechanism.

**Directed Closure and Network Structure.** Given the prevalence of directed closure in the Twitter network, one might suppose that it operates according to a relatively uniform underlying mechanism. But what we find, surprisingly, is significant heterogeneity in the amount of directed closure. We define the *closure ratio* of a  $\mu$ -celebrity  $C$  to be the fraction of  $C$ 's incoming edges that exhibit closure. If we track the closure ratio of  $C$  as edges to  $C$  are added in their temporal order, we find that the ratio stabilizes to an approximately constant value fairly early. However, the value to which the closure ratio stabilizes varies considerably from one  $\mu$ -celebrity to another, and is not closely related to the number of followers. Thus, the closure ratio appears to be an intrinsic and diverse property of users with large numbers of followers: some such users receive a clear majority of their incoming links via the closing of a directed triangle, while others receive a much smaller proportion of their links this way.

The cause of this is at some level a mystery, but to get a better understanding we look at the predictions of some basic network formation models. We present a heuristic calculation based on the preferential attachment model, suggesting that a user's closure ratio should be related to the sum of the in-degrees of the user's followers, and we find on the Twitter data that the closure ratio indeed follows this quantity more closely than simpler quantities such as the user's own number of followers. However, preferential attachment is not able to explain either the diversity of different closure ratios, or the fact that they can be large on nodes of small in-degree; to understand these effects better, we

analyze more complex models that do not incorporate copying as an overt or explicit mechanism in link formation, including preferential attachment with fitness [11] and a version of preferential attachment with embedded community structure which is related to a model of Menczer [98].

We also note that the closure ratio of a user is distinct from — and exhibits qualitatively different properties than — the *clustering coefficient* [128]. The clustering coefficient is the fraction of pairs in a node’s neighborhood that are directly linked, and in the neighborhood of a high-degree node it is almost always a small quantity, for the fundamental reason that most of a high-degree node’s neighbors don’t have enough incident edges to produce a significant clustering coefficient [125]. The closure ratio, on the other hand, is a quantity that can be quite large even for the neighborhoods of nodes with extremely large degrees.

## 4.1 Twitter Data and Micro-Celebrities

We collected a random sample of  $\mu$ -celebrities on Twitter, each with between 10,000 and 50,000 followers. For each of these  $\mu$ -celebrities  $C$ , we determine the subset of edges to  $C$  that exhibit closure.

It is an interesting fact that determining this subset does not require exact time-stamps or full network structure. Rather, it is enough to have a chronologically ordered list  $L_{in}(C)$  of the followers of  $C$ , and for each user  $A \in L_{in}(C)$ , a chronologically ordered list  $L_{out}(A)$  of the users that  $A$  follows.<sup>1</sup> From these lists, we can conclude that an edge from  $A$  to  $C$  exhibits closure if and only if there

---

<sup>1</sup>Such ordered lists were available via the Twitter API at the time we performed these analyses [63].

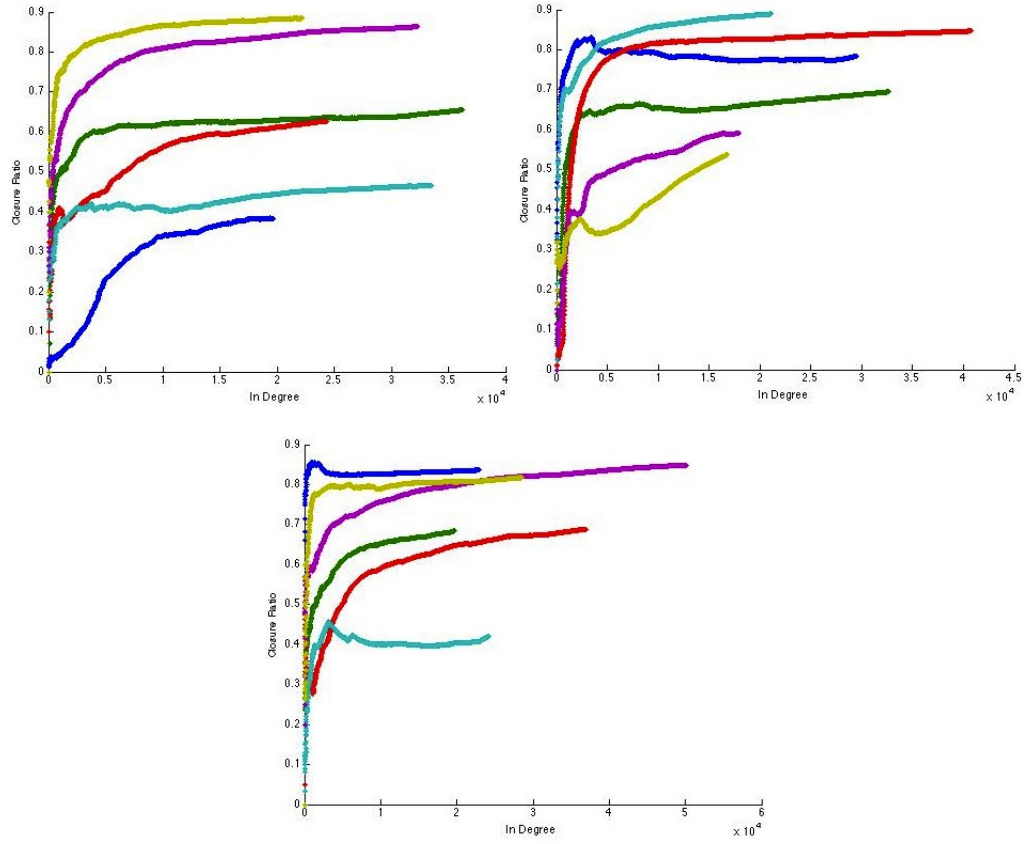


Figure 4.3: Closure ratio as a function of the arrival order of incoming edges for 18 Twitter  $\mu$ -celebrities. The following are the professions of the  $\mu$ -celebrities in each figure (from top to bottom curve). Top figure: Journalist, Venture Capital Blogger, Actor, Actor, DJ, Skateboarder. Middle figure: Comedian, Film Producer, Social Media Blogger, Musician, Actor, Journalist. Bottom figure: Comedian, TV Presenter, Actor, Musician, Film-maker, Actor.

exists a  $B$  such that  $B$  precedes  $A$  in  $L_{in}(C)$  and  $B$  precedes  $C$  in  $L_{out}(A)$ .

In Figure 4.3, we show the running fraction of edges that exhibit closure as the followers of a  $\mu$ -celebrity  $C$  arrive in chronological order. As noted in the introduction, in most cases this fraction reaches a relatively stable value quite quickly, and this stable value varies a lot from one  $\mu$ -celebrity to another. Our models in the subsequent sections will help us investigate this phenomenon.

## 4.2 Evidence for Directed Closure

We now use the randomization test described in the introduction to identify evidence for the directed closure process at work. We take the subgraph induced on the nodes in  $\{C\} \cup L_{in}(C)$ , and we insert the edges in an order selected uniformly at random from among all permutations of the edges.

Specifically, we say that a user  $A$  is  $k$ -linked to a user  $C$  if  $A$  follows  $C$ , and  $A$  also follows  $k$  followers of  $C$ . (For example, in Figure 4.2,  $A$  is 3-linked to  $C$ .) Let  $S_k(C)$  denote the set of all users who are  $k$ -linked to  $C$ , and let  $f_k$  denote the fraction of users in  $S_k(C)$  whose edge to  $C$  exhibits closure.

Now, for each  $k$  with  $|S_k| > 10$ , we approximate the expected value of  $f_k$  under the assumption that the order in which the edges are created is chosen uniformly at random. To do this, we run a simulation in which we generate a network consisting simply of a node  $A$  pointing to a node  $C$  and to  $k$  other nodes which also point to  $C$ ; we randomly choose  $|S_k|$  different orderings of the edges of this network (one corresponding to each of the  $|S_k|$  followers who are  $k$ -linked to the real  $\mu$ -celebrity); and we then determine the fraction of these random orderings in which the  $A$ - $C$  edge exhibit closure. We approximate the expected value of  $f_k$  over randomly ordered edges by the average closure ratio among 100 runs of this simulation, and we define error bars using the minimum and the maximum fraction among the 100 simulations.

We find the same trend for all the  $\mu$ -celebrities in our sample, as shown in Figure 4.4: there is some  $K$  such that for all  $k < K$  the actual value of  $f_k$  is higher than the maximum fraction from the 100 simulations. This means that at least for small values of  $k$  the fraction of edges exhibiting closure is much higher than

expected by chance. This suggests the existence of an underlying mechanism — copying of links or something producing similar observed behavior — that makes it more likely than chance to see edges that appear to be copied. For large values of  $k$ , the expected value of  $f_k$  assuming random ordering of edges becomes very large, and it is hard for the values observed in the data to lie above the error bars; we find that for large  $k$ , the actual value of  $f_k$  is very close to the average fraction among the 100 simulations and is inside the error bars.

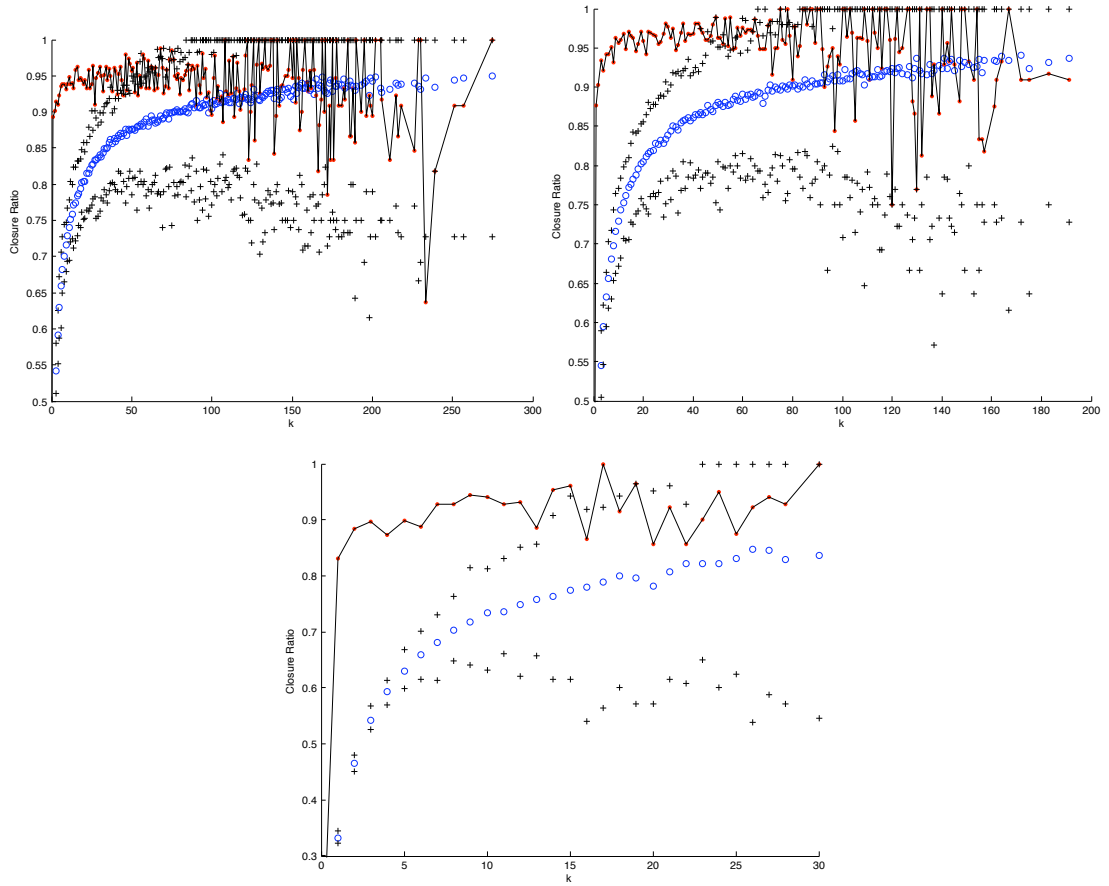


Figure 4.4: The connected dots indicate the actual value of  $f_k$ , the circles indicate the average closure ratio among the 100 simulations, and the plus signs indicate the error bars. Results for 3  $\mu$ -celebrities are shown. The trend is similar for all other  $\mu$ -celebrities

### 4.3 Preferential attachment

We would like to use probabilistic models of network formation to investigate the following two fundamental properties of directed closure in the data. First, for nodes whose in-degrees are at the level of  $\mu$ -celebrities, the closure ratio saturates to a constant  $f$  as edges arrive over time. Second, this constant  $f$  is quite different for different  $\mu$ -celebrities, and it is not closely related to the total in-degree of the  $\mu$ -celebrity.

We now compare this with the predictions of a sequence of increasingly complex models. We begin with a very basic model — a variant of the standard *preferential attachment* process, defined as follows [5, 102]:

- Fix  $\alpha \in [0, 1]$ , and  $D, N \in \mathbb{N}$ . The graph will have  $N$  nodes labeled  $0, 1, 2, \dots, N - 1$ .
- Initially (at  $t = 0$ ) the graph consists of node labeled 1 with an edge pointing to the node labeled 0.
- At each time step ( $t = j$ ) node  $j$  will join the graph with  $D$  edges directed to nodes chosen from a distribution on  $1, 2, \dots, j - 1$ . The endpoint of each edge is chosen in the following way: With probability  $\alpha$  the endpoint is chosen uniformly at random from  $\{1, 2, \dots, j - 1\}$ . With probability  $1 - \alpha$  the endpoint is chosen at random from a probability distribution which weights nodes by their current in-degree.

We run this process with different values of  $\alpha$ ,  $D$ , and  $N$  and find that preferential attachment does not achieve the desired results for  $\mu$ -celebrities. In our simulations, only nodes with very large in-degree have a reasonably large clo-



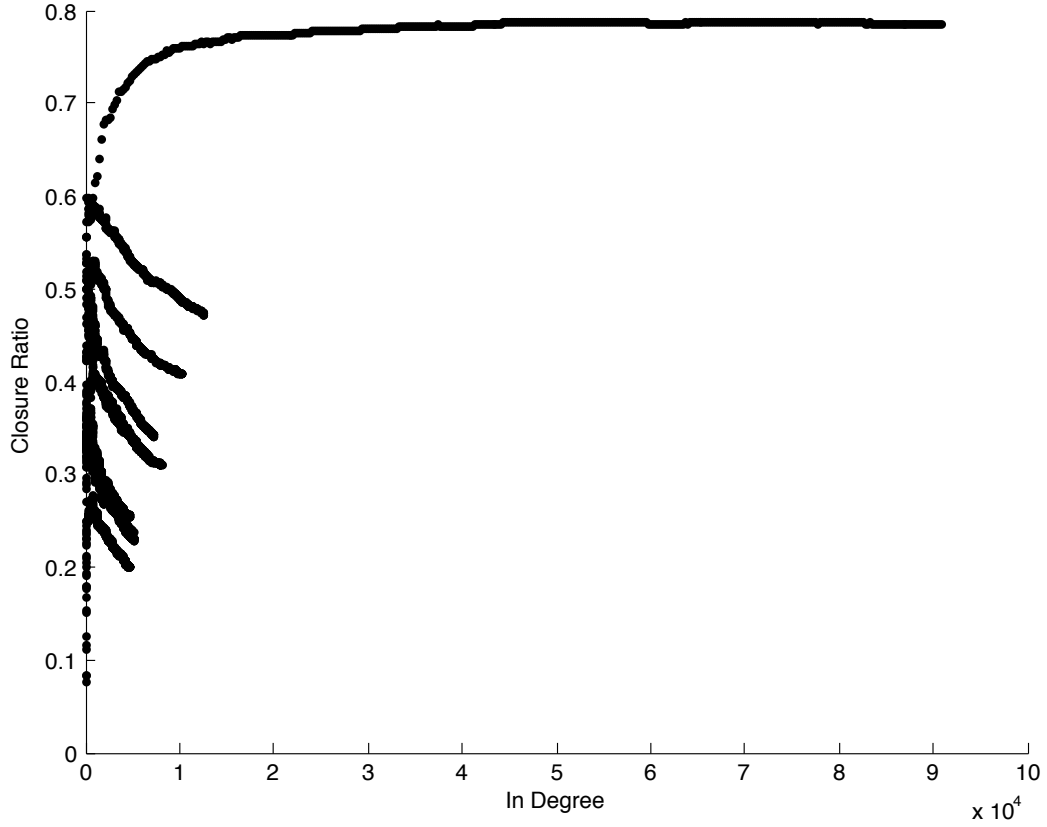


Figure 4.5: Results from the preferential attachment simulation with  $N = 200,000$ ,  $\alpha = .3$ , and  $D = 10$ . The figure shows the closure ratio as a function of edge arrival order of the 10 nodes with highest in-degree.

sure ratio, while for other nodes it is essentially zero. For those nodes with very large in-degree, the closure ratio saturates to a constant  $f$  as edges arrive, and the value of  $f$  is different for different nodes. However, the value of  $f$  is monotonically increasing as the final in-degree of node increases (See Figure 4.5).

Through a heuristic calculation we now estimate the expected closure fraction of a node in a graph generated by the preferential attachment process.

Let  $E_t$  be the total number of edges at time  $t$ ,  $N_t$  be the total number of nodes

at time  $t$ ,  $d_t(j)$  be the in-degree of node  $j$  at time  $t$ ,

$$F_t(j) = \{x : \exists e = (x, j) \text{ at time } t\},$$

$$d_t(S) = \sum_{x \in S} d_t(x), \text{ and}$$

$$S_t(j) = \alpha \frac{|F_t(j)|}{N_t} + (1 - \alpha) \frac{d_t(F_t(j))}{E_t}.$$

Note that  $S_t(j)$  is the probability that a particular edge from node  $t + 1$  is directed to a node  $k$  such that there is an edge from  $k$  to  $j$ . In other words it is the probability that an edge from node  $t + 1$  is directed to a node that points to  $j$ .

Fix a node  $j$  and an edge  $e$  coming out of node  $t + 1$ . We would like to calculate the probability of the following event  $V$ : There is another edge  $e' = (t + 1, x)$  created before  $e$  such that  $x$  points to  $j$  (i.e  $\exists$  edge  $g = (x, j)$ ). We will use  $C_{t,e}(j)$  to denote the probability of this event  $V$ . Note that we do not know which of the  $D$  edges coming out of  $t + 1$  the edge  $e$  is, or what the destination of  $e$  is. Note that if  $e$  is the first edge coming out of  $t + 1$  then the event  $V$  cannot happen; if  $e$  is the second edge coming out of  $t + 1$  then  $C_{t,e}(j) = S_t(j)$ , if  $e$  is the third edge coming out of  $t + 1$  then  $C_{t,e}(j) = [1 - (1 - S_t(j))^2]$ , and more generally if  $e$  is the  $d^{\text{th}}$  edge coming out of  $t + 1$  then  $C_{t,e}(j) = [1 - (1 - S_t(j))^{d-1}]$ . Since it is equally likely that  $e$  is any of the  $D$  edges coming out of  $t + 1$  we write

$$\begin{aligned} C_{t,e}(j) &= \frac{1}{D}[1 - (1 - S_t(j))] + \frac{1}{D}[1 - (1 - S_t(j))^2] + \\ &\quad \dots + \frac{1}{D}[1 - (1 - S_t(j))^{D-1}] \\ &= 1 - \frac{1 - (1 - S_t(j))^D}{DS_t(j)}. \end{aligned}$$

If we knew that edge  $e$  pointed to node  $j$  then the event  $V$  exactly says that  $e$  exhibits closure. Therefore if we want to know the probability that  $e$  exhibits closure given that  $e = (t + 1, j)$  we would need to calculate  $P(V|e = (t + 1, j))$ . For the

sake of our approximation, we use the unconditional probability  $P(V) = C_{t,e}(j)$  instead as our estimate of the probability that  $e$  exhibits closure. Note that the quantity  $C_{t,e}(j)$  only depends on  $j$  and  $t$ , so we define  $C_t(j) = 1 - \frac{1-(1-S_t(j))^D}{DS_t(j)}$ . In general, a given edge  $e = (x, y)$  exhibits closure with a probability of approximately  $C_{x-1}(y)$ . If  $\lim_{t \rightarrow \infty} C_t(j) = L < \infty$  then, for a large enough  $T$ , if  $t > T$  then  $C_t(j) \approx L$ . In other words, if  $t > T$  the probability that an edge coming out of node  $t$  directed to node  $j$  exhibits closure is approximately  $L$ , which in turn is approximately  $C_t(j)$ . Therefore, if  $\lim_{t \rightarrow \infty} C_t(j) = L < \infty$  and our parameter  $N$  is large enough then  $C_t(j) \approx C_{N-1}(j)$  for  $t > T$ . Hence, if  $N$  is large enough the final closure ratio of node  $j$  is approximately  $C_{N-1}(j)$ .

In Figure 4.6 we show that despite the approximations made in this argument, the calculation is a close fit to the actual closure ratios.

## 4.4 Preferential Attachment with Fitness

The fact that preferential attachment produces very few nodes with non-trivial closure ratios, and that these closure ratios are closely tied to the in-degrees, indicates the need for a more complex model. One alternative would be the use of *copying models* [74, 124], where nodes explicitly copy links from other nodes that have already joined the network. Such a mechanism builds copying into the model, generally with a tunable parameter that could be used to control quantities such as the closure ratio. However, we would like to understand whether non-trivial closure ratios — and in particular, high levels of diversity in closure ratios — can also appear in networks arising from models that do not explicitly define copying as a mechanism. As a first step in this direction,

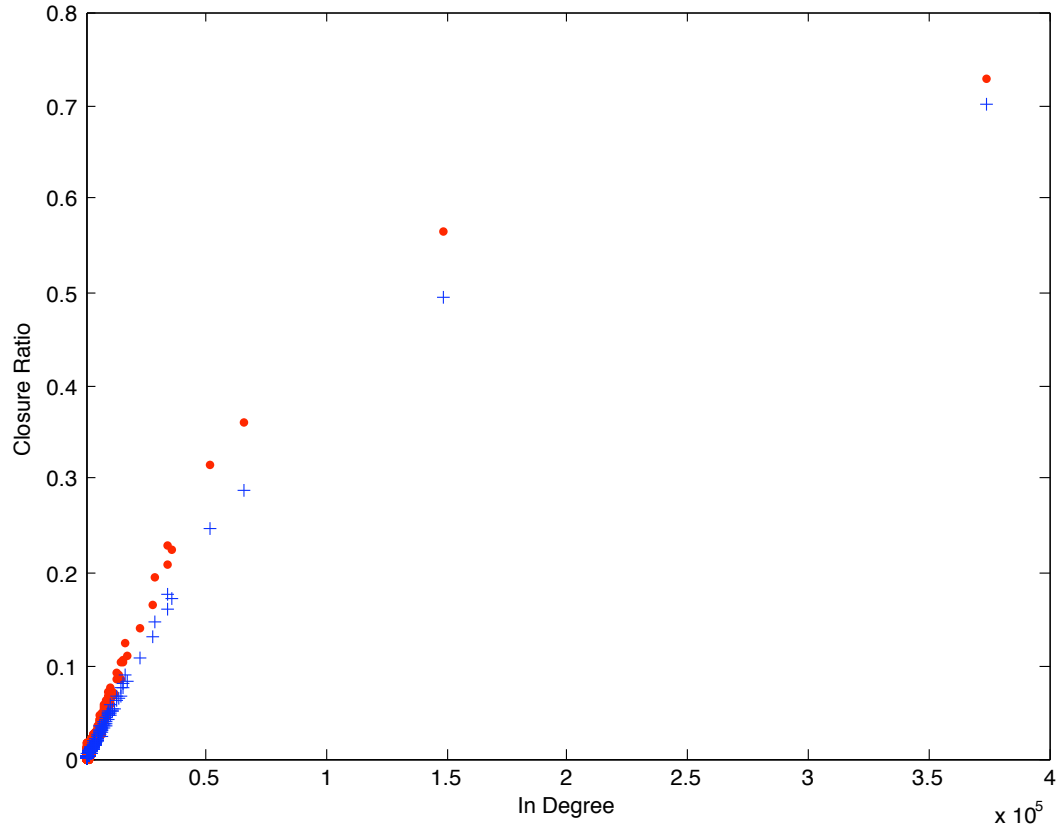


Figure 4.6: The actual closure ratio of each node  $j$  generated by the preferential attachment model with parameters  $N = 200,000$ ,  $\alpha = .3$ , and  $D = 10$  (dots) and its approximation by  $C_{N-1}(j)$  (plus signs).

we investigate an extension of preferential attachment incorporating the idea that different nodes may have different levels of inherent *fitness* or *attractiveness*, which affects how strongly they attract links [11].

Here is how this model works:

- Fix  $\alpha \in [0, 1]$ , and  $D, N \in \mathbb{N}$ . The graph will have  $N$  nodes labeled  $0, 1, 2, \dots, N - 1$ .
- Each node also has a fitness parameter  $f_i \in (0, 1)$  chosen uniformly at ran-

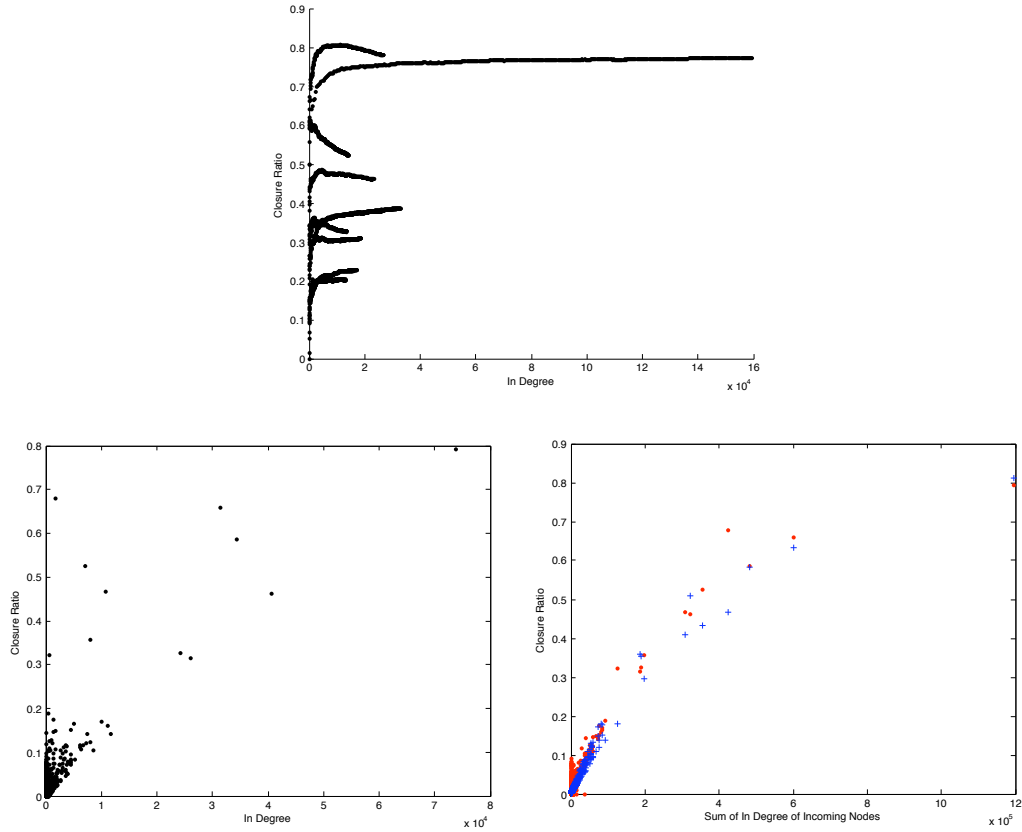


Figure 4.7: Results from the preferential attachment with fitness simulation with  $N = 200,000$ ,  $\alpha = .3$ , and  $D = 10$ . The top figure shows the closure ratio as a function of in-degree of the 10 nodes with highest in-degree. The bottom function shows the final closure ratio of each node  $j$  (dots) and its approximation by  $C_{N-1}(j)$  (plus signs).

dom.

- Initially (at  $t = 0$ ) the graph consists of node labeled 1 with an edge pointing to the node labeled 0.
- At each time step ( $t = j$ ) node  $j$  will join the graph with  $D$  edges directed to nodes chosen from a distribution on  $1, 2, \dots, j - 1$ . The endpoint of each edge is chosen in the following way: With probability  $\alpha$  the endpoint is chosen uniformly at random from  $\{1, 2, \dots, j - 1\}$ . With probability  $1 - \alpha$

the endpoint is chosen at random from a probability distribution which weights each node  $i$  by  $d_i f_i$ , where  $d_i$  is the node's current in-degree.

We run simulations of preferential attachment with fitness, with different parameters, and find an improvement from the simple preferential attachment model. A node's final closure ratio is not correlated with the final in-degree of the node, which matches what we found in our data set. However, just like in the simple preferential attachment model, very few nodes have a closure fraction that is non-trivially larger than 0 (see Figure 4.7). In particular, for the nodes that would correspond to  $\mu$ -celebrities, the fraction is basically zero. This is not consistent with the data, which shows that  $\mu$ -celebrities can have very large closure ratios.

We find that the heuristic calculation for the closure ratio we derived for the preferential attachment model is very accurate for preferential attachment with fitness as well. Furthermore, from the calculation we see that for a node  $j$  the term  $d_t(F_{N-1}(j))$  (the sum of the in-degree of nodes that point to  $j$ ) is the most important in determining the closure ratio when  $\alpha$  is small. For preferential attachment with fitness, the closure ratio of a node  $j$  is much more correlated with  $d_t(F_{N-1}(j))$  than with the in-degree of  $j$  (see Figure 4.7). This is also the case for the  $\mu$ -celebrities in our data set (see Figures 4.8 and 4.9), which means that in determining a user's closure ratio, the more important variable seems to be not the number of followers the user has but the total number of followers of those who follow the user.

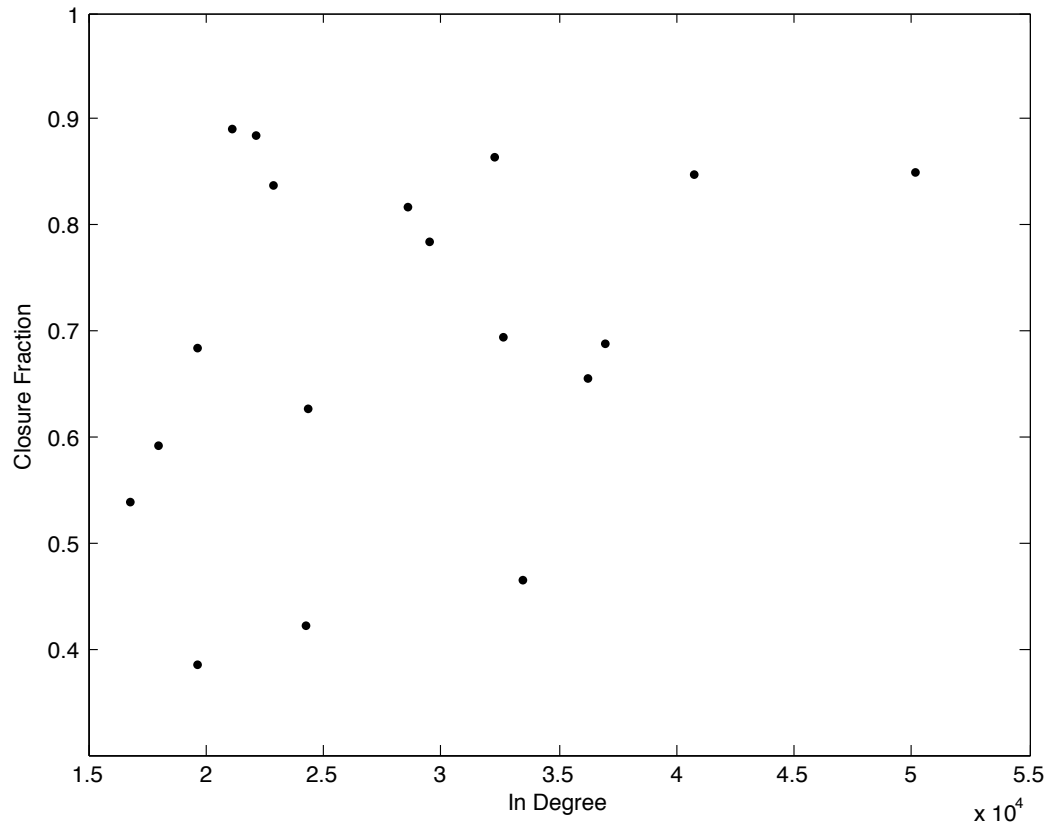


Figure 4.8: Closure ratio as a function of In-Degree.

## 4.5 Preferential Attachment with Communities

The previous model, incorporating fitness, manages to produce heterogeneity in the closure ratios, but it still only produces very few nodes for which the closure ratios are non-trivial. We now present a model in which many nodes will have non-trivial closure ratios.

The model is *preferential attachment with communities*: we assume that each node belongs to a particular community of nodes, and the node is more likely to attach to nodes from its own community than to nodes from other communities.

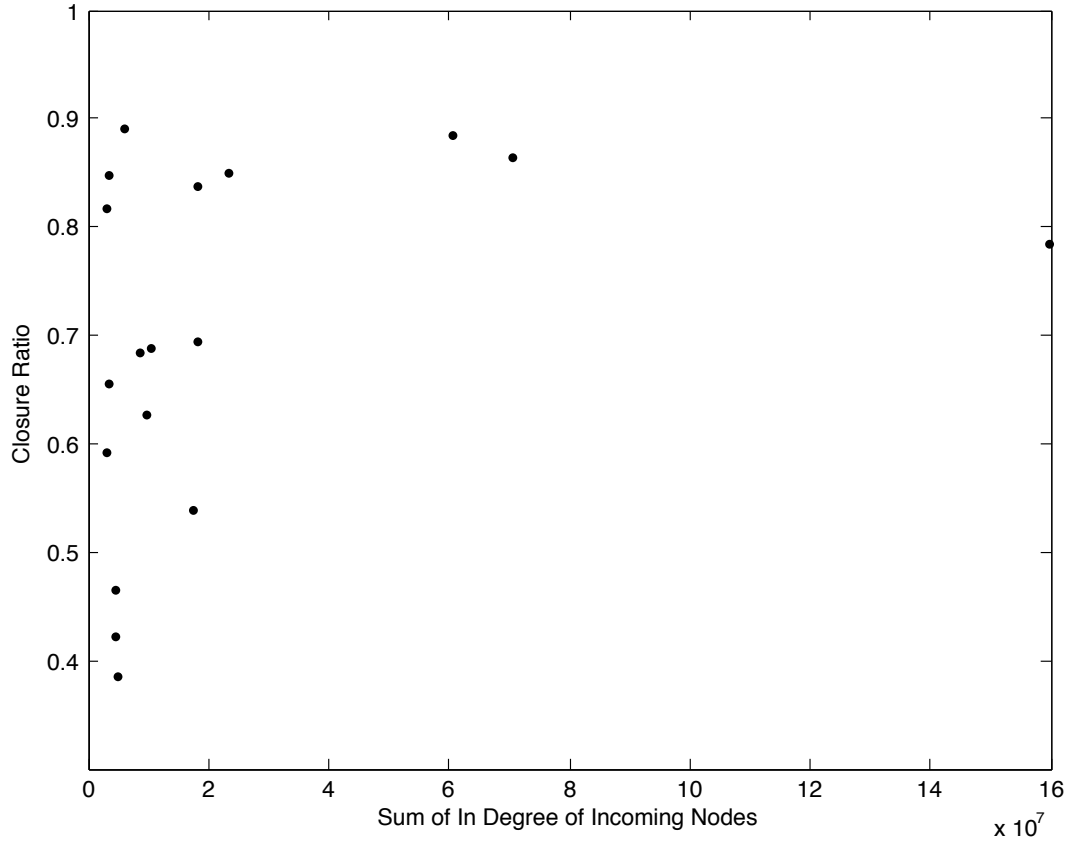


Figure 4.9: Closure ratio as a function of the Sum of In-Degree of Incoming Nodes.

Specifically:

- Fix  $\alpha \in [0, 1]$ ,  $\beta \in [.5, 1]$ , and  $C, D$ , and  $N \in \mathbb{N}$ . The graph will have  $N$  nodes labeled  $0, 1, 2, \dots, N - 1$  and there will be  $C$  communities.
- Initially (at  $t = 0$ ) the graph consists of the  $C$  communities, each with two nodes, one pointing at the other.
- At each time step ( $t = j$ ) node  $j$  will join the graph and will be assigned a community uniformly at random. Then  $j$  will create  $D$  edges directed to nodes chosen from a distribution on  $1, 2, \dots, j - 1$ . The endpoint of each edge



is chosen in the following way: With probability  $\beta$  the endpoint will be a node from the same community as  $j$ ; with probability  $1 - \beta$  the endpoint will be chosen from any of  $1, 2, \dots, j - 1$ . With probability  $\alpha$  the endpoint will be chosen preferentially (i.e. at random from a probability distribution which weights nodes by their current in-degree) and with probability  $1 - \alpha$  the endpoint will be chosen uniformly at random from the set of nodes already determined.

Simulations with different parameters show that this model generates nodes whose closure ratios converge as in-degree increases (see Figure 4.10), and the final fraction is not closely related to the in-degree as it was in the case of simple preferential attachment. Furthermore, the nodes that would correspond to a  $\mu$ -celebrity level of in-degree can have reasonably large closure ratios.

It is also interesting to note that the sum of a node's followers' in-degrees, an important parameter in the previous two models, still plays a role here, but with a twist: as Figure 4.11 shows, a node's closure ratio is more closely correlated with the sum of in-degrees of the followers *from its own community* than with the sum of the in-degrees of all its followers. It would be interesting to explore this quantity on the Twitter data, using different approximations of community structure in Twitter.

## 4.6 Discussion

We have studied the process of directed closure in information networks, developing a definition and methodology for evaluating it, and providing evidence for directed closure in the follower network of Twitter. We also found that the

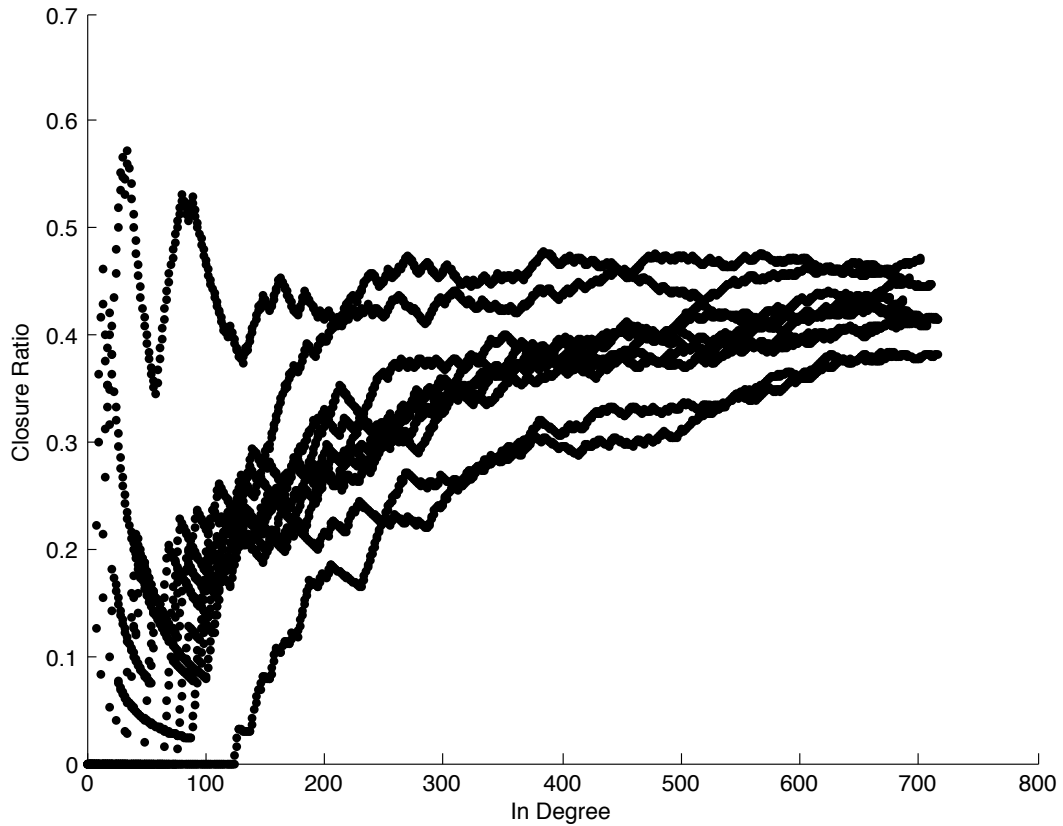


Figure 4.10: The closure ratio as a function of in-degree for the 10 nodes with highest in-degree. Preferential attachment with communities simulation with  $N = 200,000$ ,  $\alpha = .3$ ,  $\beta = .8$ ,  $C = 1,000$ , and  $D = 10$ .

extent of directed closure varies considerably between the sets of followers of different popular users. A sequence of models generalizing the principle of preferential attachment provide some explanation for our findings, and identify a more subtle parameter — the sum of the in-degrees of one’s followers — that is related to the extent of directed closure.

It is an interesting direction for further work to try understanding better the causes of heterogeneity in the closure ratios of micro-celebrities on Twitter, and the extent to which identifying communities in the Twitter network structure

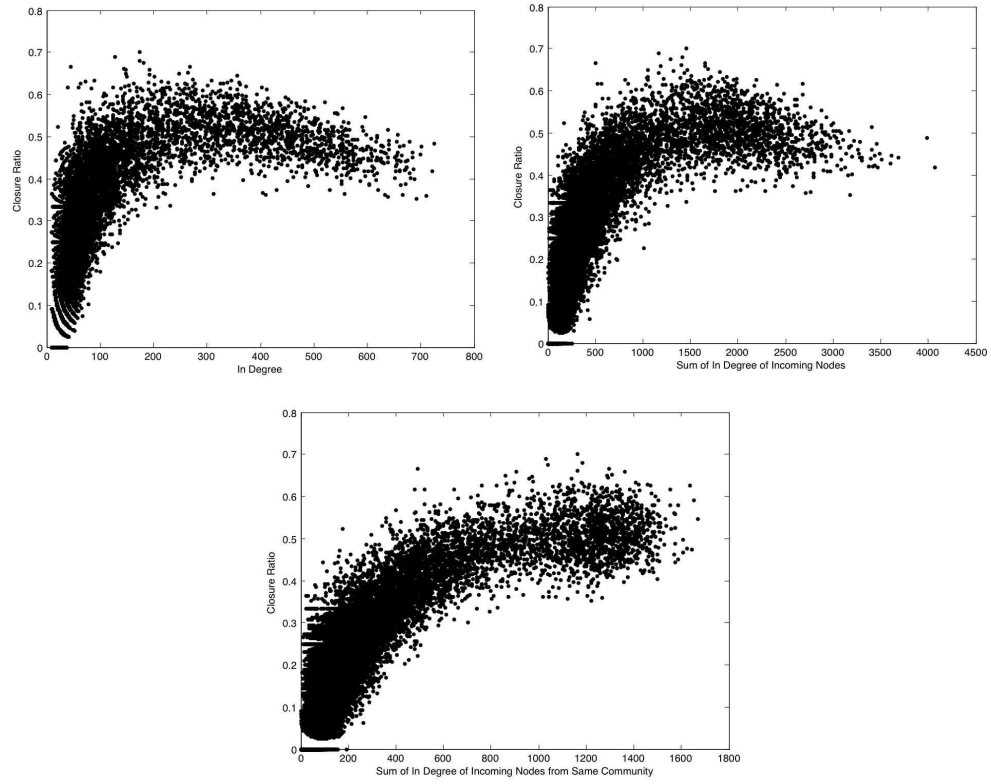


Figure 4.11: Results from the preferential attachment with communities simulation with  $N = 200,000$ ,  $\alpha = .3$ ,  $\beta = .8$ ,  $C = 1000$ , and  $D = 10$ .

can help evaluate the more detailed predictions of preferential attachment with communities. It will also be interesting to explore comparative analyses of these measures on other information networks.

## CHAPTER 5

### MAINTAINING TIES IN SOCIAL MEDIA

In studying the interactions on a social media site, a basic question is to understand what causes relationships among users to be strengthened and what causes them to weaken. This is an issue that is not well understood: there are multiple forces that govern the strengths of social ties and pull in competing directions. It is an important problem to design methods of analysis for these systems that can begin to separate out the effects of these different forces. Existing work in on-line domains has approached this issue by identifying dimensions that characterize the strength of ties [36], and by incorporating factors such as triadic and focal closure [71], similarity among individuals [6, 27, 7, 72], and the role of positive and negative relationships [84].

In this chapter, we develop an analysis framework through which we can use data from social media sites to begin isolating the effects of three distinct social forces on the strengths of relationships: *balance*, *exchange*, and *betweenness*. We begin by describing how these forces operate in a social media context, which will also make clear the sense in which they can produce opposite effects. For this discussion we will focus on undirected links, in which relationships are symmetric.

**Balance and Exchange.** First, we consider the force of balance. Suppose we have a user  $B$  who is friends with users  $A$  and  $C$ . The principle of balance argues that if  $A$  and  $C$  do not have a social tie, this absence introduces latent strain into the  $B$ - $A$  and  $B$ - $C$  relationships, and this strain can be alleviated if an  $A$ - $C$  tie forms [54, 107]. Hence, balance is a force that causes the formation of an  $A$ - $C$  tie

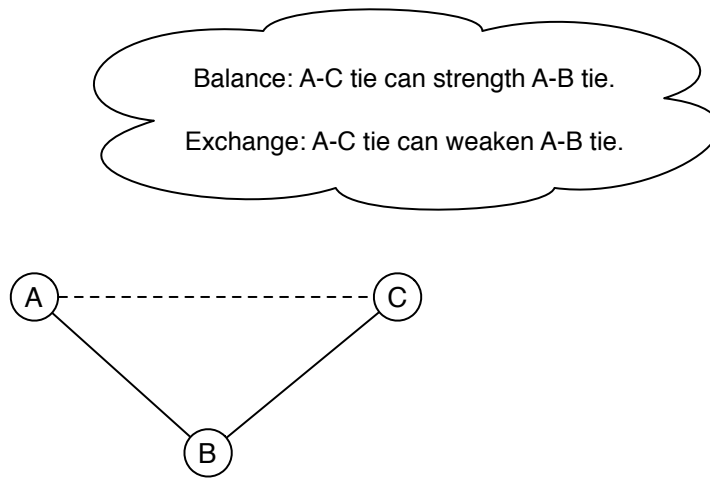


Figure 5.1: The theories of balance and exchange postulate the effect of A and C forming a relationship on the B-A and B-C relationships.

to strengthen the *B-A* tie, when *C* is also linked to *B*.<sup>1</sup>

Counterbalancing this is an equally natural force, which is the principle of *exchange* [33, 130]. Let's return to the user *B* who is friends with users *A* and *C*. If *A* were to become friends with *C*, this provides *A* with more social interaction options than she had previously. The theory of exchange argues that this makes *A* less dependent on *B* for social interaction, thereby weakening the *B-A* tie.

Figure 5.1 is a schematic diagram of the forces of balance and exchange as they act on a set of three nodes. Our first set of analyses studies the aggregate effect of these forces on the communication patterns between Twitter users. For this, we say that a *tie* between two Twitter users has formed when they have

<sup>1</sup>One sees balance theory applied in two related contexts when we consider scenarios such as this, when *B* has positive relations with *A* and *C*. In one line of argument, the absence of an *A-C* link produces stress that needs to be resolved. A related line of argument considers situations in which there is in fact antagonism between *A* and *C*, which produces even stronger forms of stress [20]. Both of these situations point to the same conclusions, and both fall under the principle of balance.

each sent at least 3 @-messages to the other.<sup>2</sup> We examine ties between users in a large collection of public tweets. We also consider scenarios, such as the one pictured in Figure 5.1, in which a user  $B$  has ties to users  $A$  and  $C$ , and look at cases in which an  $A$ - $C$  tie does or does not form.

**Decaying Relationships and Outside Opportunities.** We find first of all that the formation of an  $A$ - $C$  link in our Twitter data makes it significantly more likely that the  $A$ - $B$  tie will persist (as measured by the generation of future messages from  $A$  to  $B$ ). At one level, this points to the dominance of balance over exchange in this particular scenario; however, as we investigate the effect of tie formation on tie persistence more closely, a more subtle picture emerges. Going back to users  $A$ ,  $B$ , and  $C$ , suppose that we consider the effect on the  $A$ - $B$  tie of  $A$ 's sending  $k$  messages to arbitrary users other than  $B$ , for some relatively large value of  $k$  — potentially even requiring these messages to go to users not linked to  $B$ . Even in this case, these messages from  $A$  to others lead to an increase in the persistence of the  $A$ - $B$  tie.

This observation underscores the need to be careful in reasoning about how the persistence of ties operates on a social media site. One might suppose, via the principle of exchange, that the  $k$  messages from  $A$  to others divert  $A$ 's attention from  $B$ , to the detriment of the  $A$ - $B$  tie. But we should step back and think about the full set of activities that might draw  $A$  away from  $B$ . Interaction with other users on Twitter is one source of such activities. However, there are many activities completely outside Twitter that might draw  $A$ 's attention away from

---

<sup>2</sup>@-messages are a basic Twitter mechanism in which one user directs a tweet to another; since they are used between people who know one another as well from users toward celebrities, we require multiple reciprocations before we consider the messaging to constitute evidence of a tie.

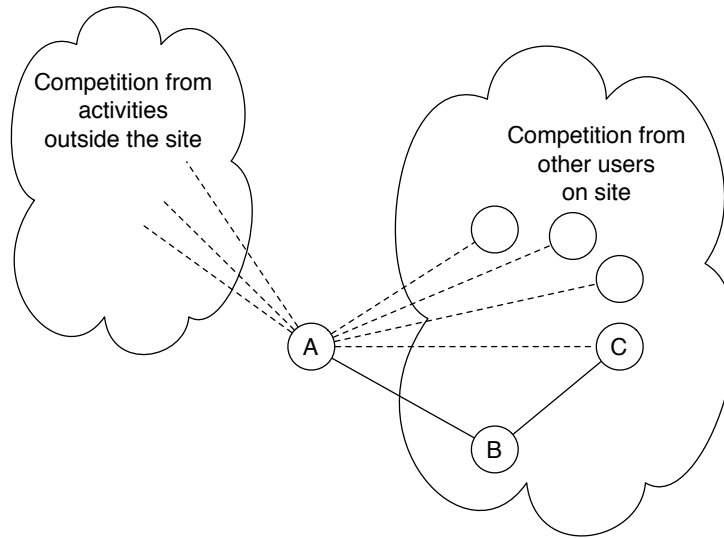


Figure 5.2: Outside influence: The A-B relationship is potentially weakened not only by additional relationships within the online social network, but also by activities that altogether draw users away from the network.

*B* as well. Thus, abstractly, the picture from Figure 5.1 should be expanded to look more like the larger picture in Figure 5.2.

In context of Figure 5.2, the principle of exchange is not irrelevant to the discussion, but we are applying it too narrowly if we view other Twitter users as the only sources of outside opportunities for *A* in the *A-B* relationship. And the point, then, is that  $k$  messages from *A* to many users other than *B* still provide strong evidence that *A* is actively involved in Twitter, rather than in other activities. This increased involvement makes it easier for *A*'s Twitter activity to “spill over” to the *A-B* tie.

In Section 5.3, we consider ways of capturing this spillover effect, and propose a reconceptualization of exchange theory in the particular context of social media to integrate the outside opportunities of a user *A* at both the “micro” level

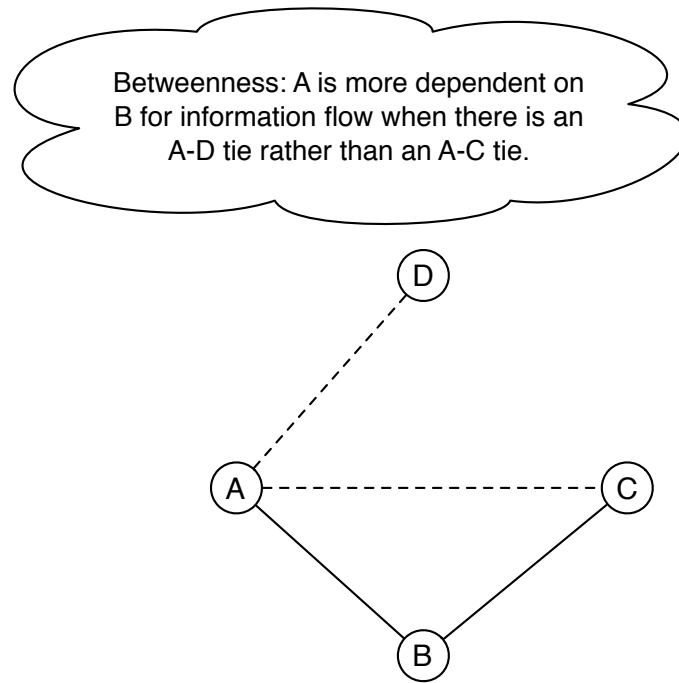


Figure 5.3: Betweenness postulates that A is more dependent on B for information flow when A connects to nodes that are not connected to B than when she connects to nodes connected to B.

(to other users on the site) and the “macro” level (to potentially unobserved activities off the site).

This framework also suggests an important methodological consideration that is underscored by our analyses. Social media sites are domains in which the typical relationship exists in a state of rapid decay, since either user involved in the relationship may begin to rapidly reduce their involvement in the site, or leave it altogether and never return. Such issues are much less of a constraint (even if they are present at lower levels) in analyses of relationships in the physical world — but in on-line settings, they need to be carefully controlled for.



**Balance and Betweenness.** Given these considerations, we explore a further set of questions about social forces and relationships in which we control for  $A$ 's overall level of involvement in the site. Specifically, consider again a user  $B$  who has ties with users  $A$  and  $C$ . Now, let a fixed amount of time pass, and consider two possible scenarios: (i)  $A$  forms a tie with  $C$ , or (ii)  $A$  forms a tie with a user  $D$  who is not connected with  $B$ . In which scenario is the  $A$ - $B$  more persistent? (See Figure 5.3.) Both (i) and (ii) provide evidence of comparable involvement by  $A$  in the site, and so we must look to the finer structure of the interaction pattern to decide which has a more positive effect on the  $A$ - $B$  tie.

As before, the principle of balance argues that the  $A$ - $B$  tie should be more strengthened in scenario (i). The principle of exchange is a bit tricky to apply here, but we can use the principle of *betweenness* instead to identify a natural argument that says that scenario (ii) should be better for the  $A$ - $B$  link. The principle of *betweenness* is used, for example, by Burt [18] in his formulation of the theory of structural holes.

The argument for betweenness is as follows. Twitter is an environment in which access to information, and the flow of information, is a crucial force in the shaping of users' activities — consider, for example, the set of social and informational links that are activated whenever a piece of content is extensively retweeted (repeated by users). As a result, when there is no  $A$ - $C$  link, user  $B$  plays an important brokerage role in her relationship with  $A$ :  $B$  provides  $A$  with access to information from  $C$ . If a direct  $A$ - $C$  tie forms, this brokerage role is sharply diminished; on the other hand, the role is not as strongly diminished if  $A$  forms a tie with  $D$ . Thus, considerations of betweenness and brokerage suggest that the  $A$ - $B$  might persist more strongly in scenario (ii), with the formation of

an  $A$ - $D$  tie, rather than in scenario (i), with the formation of an  $A$ - $C$  tie.

In Section 5.2, we carry out a careful analysis of this trade-off, finding significant evidence that the balance argument is operating more strongly than the betweenness argument in the setting of Twitter: the closing of the  $A$ - $B$ - $C$  triangle (as in scenario (i)) has a more positive effect on the  $A$ - $B$  relationship than the formation of ties by  $A$  that leave it open (as in scenario (ii)).

**Persistence of Ties.** Finally, in Section 5.4, we develop further methodologies for analyzing the persistence of relationships in social media domains such as Twitter, given the rapid rate at which they decay over time. In particular, we identify fundamental asymmetries in the way that relationships ramp up in intensity compared to the way in which they fall off after their peak level of activity, and we show how the closing of triads in the vicinity of a tie can have important effects on its persistence.

## 5.1 Data Set and Network Extraction

We have collected and processed a large corpus of data from the Twitter social network. From August 2009 until January 2010, we crawled Twitter using their publicly available API. Twitter provides access to only a limited history of tweets through its search mechanism; however, because user identifiers have been assigned contiguously since an early point in time, we simply crawled each user in a comprehensive range. Due to limitations of the API, if a user has more than 3,200 tweets we can only recover the last 3,200 tweets; all messages of any user with fewer than this many tweets are available. We collected over

three-billion messages from more than 60 million users during this crawl.

The primary analysis of this data is to extract all @-messages and build a temporal network of ‘attention relationships.’ A directed edge exists from user  $A$  to  $B$  if  $A$  sends at least  $k$  @-messages to  $B$ ; the time this edge is created,  $t_D(A, B)$ , is the time at which the  $k$ th @-message is sent. In our analyses we use  $k = 3$ . There are multiple ways of defining a network, and our definition is one way of defining a proxy for the attention that a user  $A$  pays to other users. The resulting network contains 8,509,140 non-isolated nodes and 50,814,366 links.

From this directed, temporal network we extract an undirected, temporal network of ties. An undirected edge between two users  $A$  and  $B$  is formed when  $A$  has sent at least 3 @-messages to  $B$  and  $B$  has sent at least 3 @-messages to  $A$ . The edge  $E = (A, B)$  has time-stamp equal to  $t(A, B) = \max\{t_D(A, B), t_D(B, A)\}$ , the later of the times when the two directed edges were formed. This tie network contains 20,492,393 ties between 3,701,860 users, and although fewer than half of the users remain in the tie network, over 80% of attention relationships contribute to a tie.

We define an *open triad*  $O$  as a graph of three nodes  $A$ ,  $B$ , and  $C$  containing the ties  $(A, B)$  and  $(B, C)$ . The time-stamp of the open triad is  $O_t = \max\{t(A, B), t(B, C)\}$ , the time at which the last of the two ties forms. Open triads  $O = (A, B, C)$  in which the undirected  $(A, C)$  edge eventually forms are said to *close*. We define an open triad that closes  $d$  days after  $O_t$  ( $t(A, C)$  is  $d$  days after  $O_t$ ) to be a *d-closed triad*.

## 5.2 Balance Vs. Betweenness

We begin by considering the contrast between balance and betweenness discussed in the introduction. We take an open triad  $(A, B, C)$ , and as in Figure 5.3, we compare the amount of interaction from  $A$  to  $B$  after one of the following two events takes place: (i) the  $A$ - $C$  tie forms, or (ii)  $A$  forms a tie with a user  $D$  who is not connected to  $B$ . Because we have recorded not only the evolution of a triad (whether it closed or not), but also the communication times, we can control for factors such as the delay between triad formation and the creation of the additional tie. Additionally, we will control for  $A$  being ‘active’; we make sure that  $A$  was communicating when the triad formed, when the new tie forms, and some time after the new tie formed. In this way, we will not end up studying phenomena that arise primarily because users are immediately leaving the site.

**Representing the competing scenarios.** In particular, we consider the percentage of messages that  $A$  directs to  $B$  in two comparison sets of triads designed to represent scenarios (i) and (ii). First, we choose a value for  $d$  and consider all  $d$ -closed triads; we also want to guarantee that  $A$  had a certain minimum level of activity overall, so we require that  $A$  sent between 200 and 1000 messages in total after the open triad  $(A, B, C)$  was formed, and moreover that  $A$  sent at least one message 1,  $d$ , and  $2d$  days after the open triad was created. This subset of  $d$ -closed triads with these conditions ensuring  $A$  is sufficiently active forms our population for scenario (i).

For scenario (ii), we want an open triad  $(A, B, C)$  where  $A$  sends a message to a node not connected to  $B$ . Thus, for each triad  $O' = (A, B, C)$  that never closes, we look at all of the nodes  $D$  that are not connected to  $B$ , and with which  $A$  forms

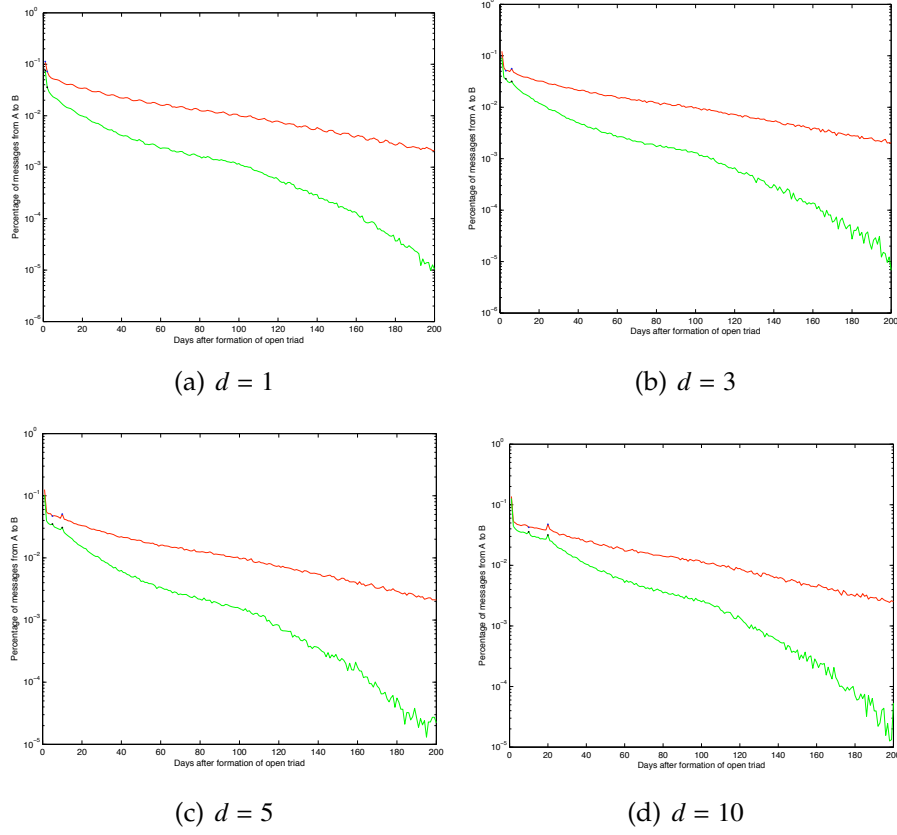


Figure 5.4: Percentage of message from  $A$  to  $B$  vs. the number of day after creation of open triad. The green curve is based on the  $d$ -open triads and the red curve is based on the  $d$ -closed triads.  $A$  must have sent from 200 to 1000 messages in total after day  $= 0$  and  $A$  must have sent at least one messages on days  $1, d$ , and  $2d$ .

a tie after  $O'_i$ . We pick such a node  $D$  at random and say that  $O'$  is  $d$ -open, where  $d$  is the number of days after  $O'_i$  that the  $A$ - $D$  tie formed. As before, we also require that  $A$  sent between 200 and 1000 messages after the open triad  $(A, B, C)$  was formed, and that  $A$  sent at least one message  $1, d$ , and  $2d$  days after the open triad was created. This population of  $d$ -open triads with these conditions on  $A$  forms our population for scenario (ii).

For each population, we measure the percentage of  $A$ 's communication that goes toward  $B$ , as a function of the time since the formation of the open triad.

As noted in the introduction, relationships on social media sites have a default tendency to decay, but by observing which scenario provides a slower aggregate decay rate for the  $A$ - $B$  tie, we can begin to learn about the different effects of balance (scenario (i)) and betweenness (scenario (ii)).

**Results.** In Figure 5.4, we adopt this test with  $d$  chosen to be 1, 3, 5, and 10 days. Each plot shows the average percentage of messages that  $A$  sent to node  $B$  as a function of the number of days after  $O_t$ . The red curve is based on the  $d$ -closed triads, while the green curve is based on the  $d$ -open triads.

We observe first that for all choices of  $d$ , the red curve decreases at a slower rate than the green curve. This indicates that the  $A$ - $B$  tie decays more slowly in the population corresponding to scenario (i). But beyond this, the gap between the two curves is widening: the rate at which they decrease is separating. After day 100 (about three months after the formation of  $A$ 's additional connection), the communication percentage for the open triads decreases at a noticeably faster rate. This suggests that closing the triad benefits communication from  $A$  to  $B$  by slowing the inevitably decreasing amount of online interaction.

In interpreting these results as evidence for the effect of balance, it is important to understand that the formation of the  $A$ - $C$  tie is not causing the extent of  $A$ - $B$  interaction to increase in an absolute sense, but rather for its rate of decay to be slowed. In general, the effect of social forces on relationships in our analysis is ubiquitously modulated by the overall rate of link decay on Twitter.

### 5.3 Exchange Theory and Spill-Over Effects

In the previous section, we observed that in the triad  $(A, B, C)$  the communication between  $A$  and  $B$  benefits in the long run from the triad's closing. At a more general level, we will now ask what can be predicted about the  $A$ - $B$  interaction from knowledge of how active  $A$  was with respect to users other than  $B$ .

Exchange theory posits that as  $A$  has more “outside options” provided by communication partners who are not  $B$ ,  $A$  will spend less time communicating with  $B$ . One hypothesis, then, is that as  $A$  spends more time talking to her friends who are not  $B$ ,  $A$ 's communication with  $B$  will decrease. Alternatively, we can consider a simple model based on the schematic picture in Figure 5.2, where  $A$  first decides how much time to spend on Twitter, and then divides that time evenly between all of her friends on Twitter. According to this model, the more time  $A$  spends talking to anyone on Twitter, the more time she will spend talking to  $B$  as well.

We test these two predictions by plotting the number of messages  $A$  sends to everyone but  $B$  vs. the number of messages that  $A$  sends to  $B$  for various points in time after the creation of the  $A$ - $B$  edge. The plots have the same general shape up to several weeks after the creation of the edge. In Figure 5.5, we present the plot for three days after the creation of the edge. The figure shows a pattern of monotonic increase, which suggests that the second model is a better approximation to the real outcome: the more  $A$  talks to anyone on Twitter, the more she talks to  $B$  as well.

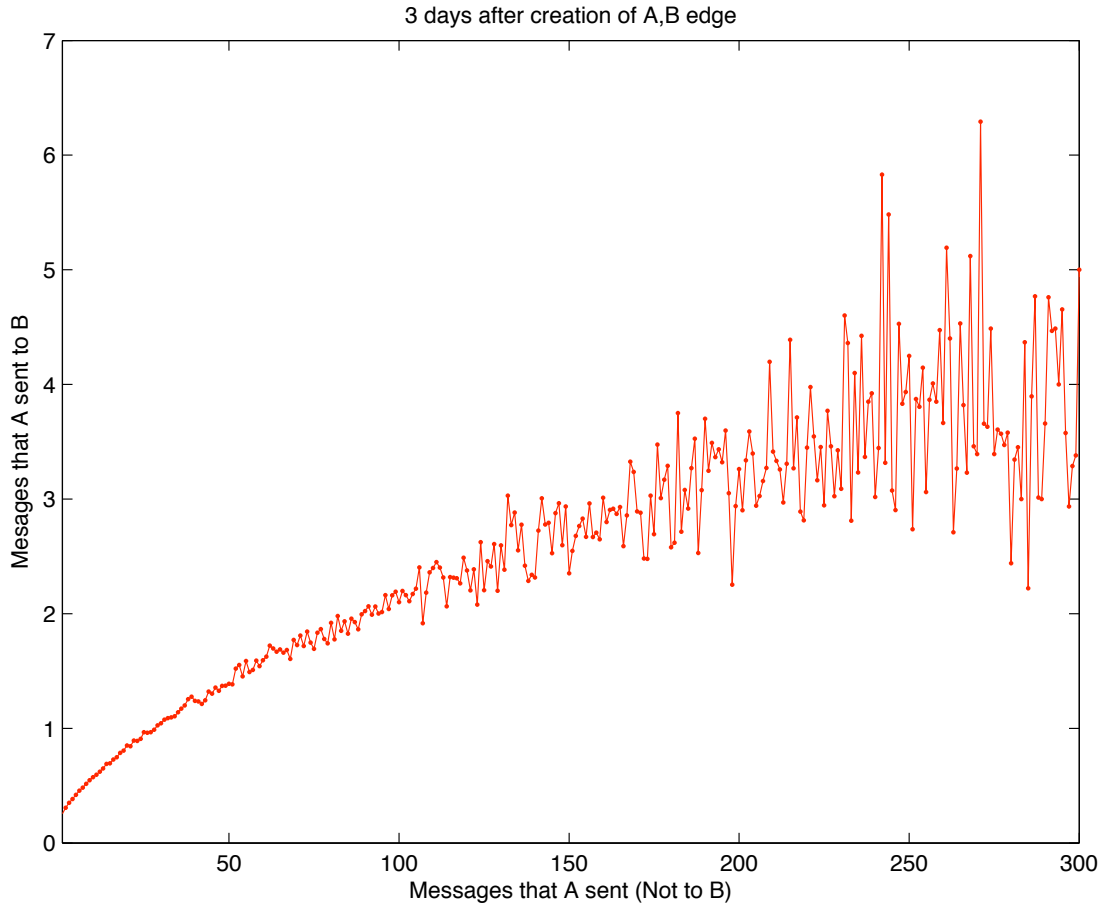


Figure 5.5: Number of messages  $A$  sends to everyone but  $B$  vs. number of messages  $A$  sends to  $B$ , 3 days after the creation of the  $A$ - $B$  edge.

**The Role of Balance in Spill-Over Effects.** This analysis makes precise the sense in which we think of  $A$ 's activity toward users other than  $B$  as “spilling over” in a positive way toward  $B$ . We now show that the principle of balance can enhance this spill-over effect. To do this, we consider the set-up above, but vary the number of  $A$ 's messages that go to users with whom  $B$  also has ties.

In particular, Figure 5.6 depicts the following analysis. We consider the messages sent by  $A$  to users other than  $B$ , and ask what fraction of these messages go to users  $C$  with whom  $B$  also has a tie. What Figure 5.6 shows is that the per-



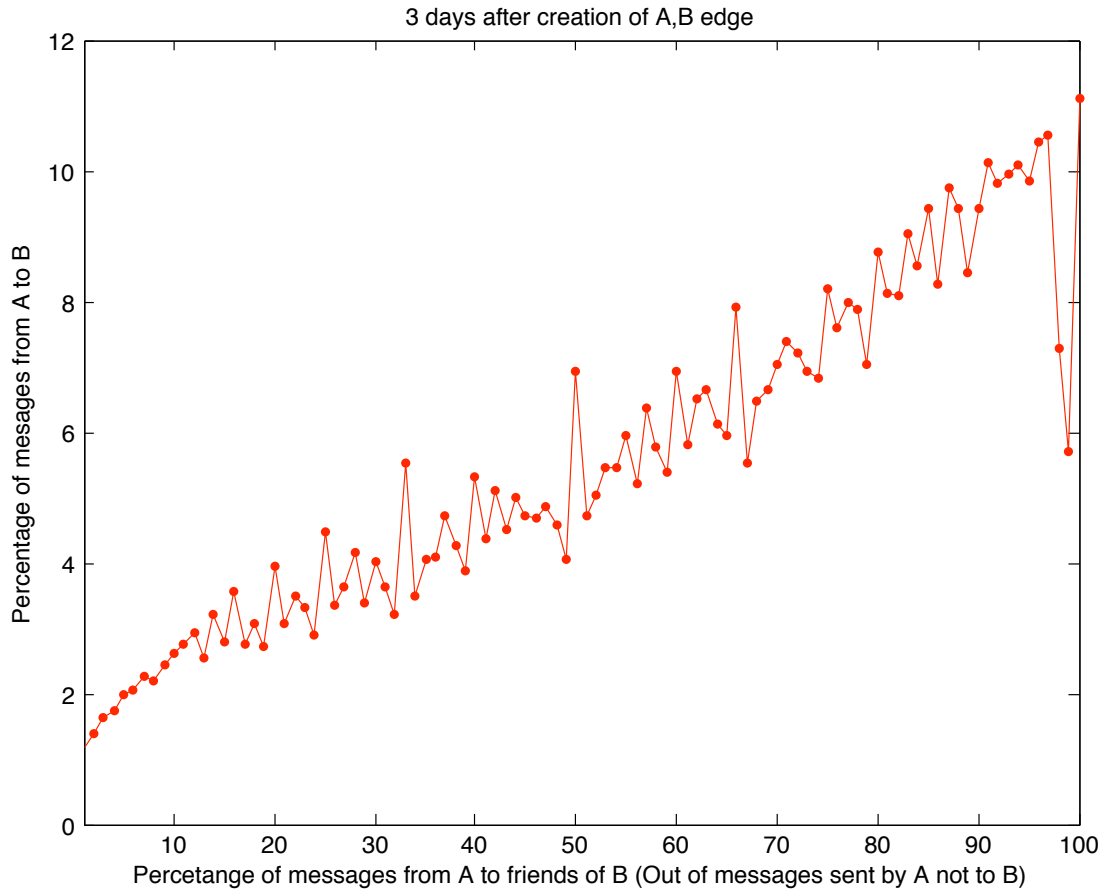


Figure 5.6: Percentage of messages that  $A$  sends to  $B$  as a function of the percentage of  $A$ 's non- $B$  messages that go to friends of  $B$ . These messages take place 3 days after the creation of the  $A$ - $B$  edge.

centage of messages from  $A$  to  $B$  increases as the percentage of messages from  $A$  to  $B$ 's friends increases: in other words, the spill-over in  $A$ 's activity toward  $B$  is accentuated when  $A$ 's activity toward users other than  $B$  takes place with friends of  $B$ .

There is a respect in which Figure 5.6 can be a bit subtle to interpret, based on the fact that it aggregates many users  $A$  of different activity levels. As a result, we show (in Figure 5.7) a related analysis in which the set-up is identical except that we require  $A$  to have sent exactly 10 messages to users other than

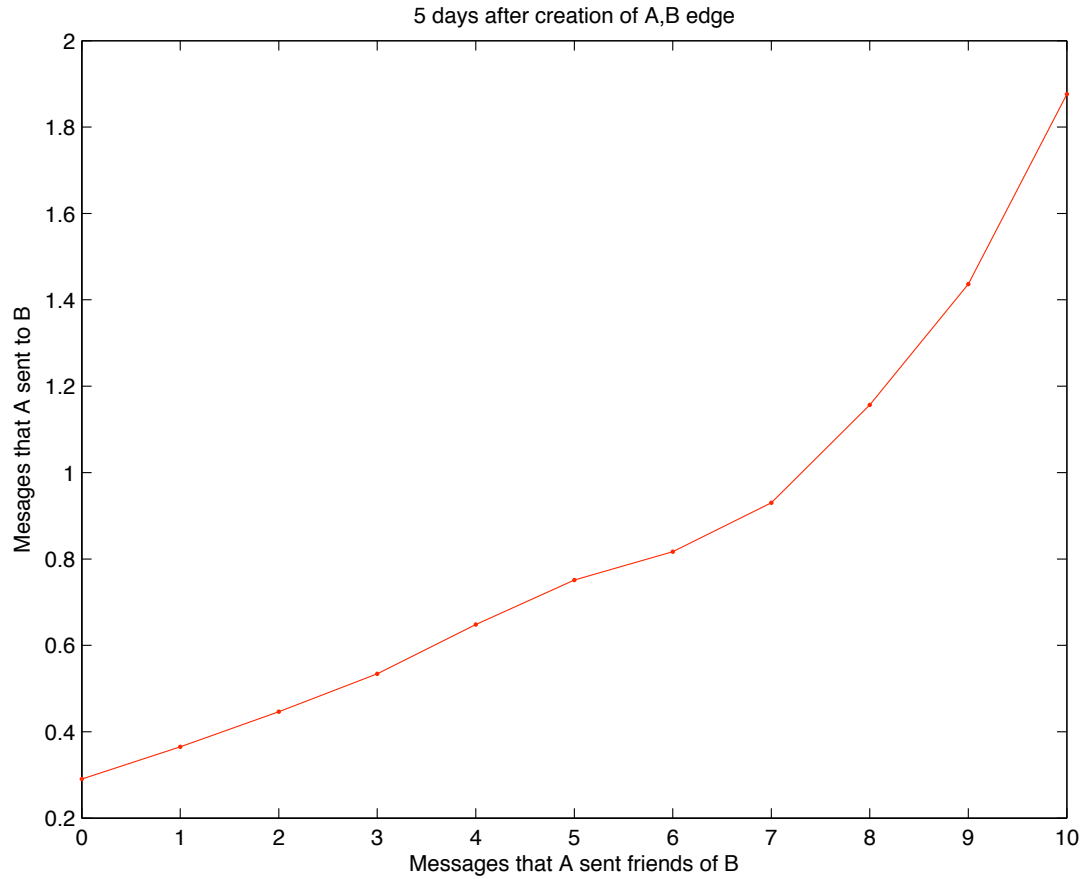


Figure 5.7: Number of messages  $A$  sends to friends of  $B$  vs. number of messages  $A$  sends to  $B$ , five days after the creation of the  $A$ - $B$  edge. Node  $A$  sent exactly 10 messages in total to users other than  $B$ .

$B$ . We then ask: how many messages does  $A$  send to  $B$ , as a function of the number (out of 10) of these non- $B$  messages that go to friends of  $B$ ? Again we find that the spill-over in  $A$ 's activity is enhanced when  $A$ 's non- $B$  activities include many friends of  $B$ . Indeed, we see in Figure 5.7 a striking super-linear relationship whereby the spill-over effect ramps up very rapidly once most of  $A$ 's non- $B$  communication is directed at friends of  $B$ .

**A Situation with Apparent Lack of Spill-Over.** Thus far we have not seen any situations in which  $A$ 's activity toward users other than  $B$  has had any kind of negative effect on the  $A$ - $B$  tie. Here we identify the possibility of one such situation, leaving the underlying mechanism for it as an open question.

The situation is the following. Figure 5.8 zooms in around the days  $d$  and  $2d$  (in this case 10 and 20) on the curves from Figure 5.4. We observe that the green curve has jumps on days 10 and 20, while the red curve only has a jump on day 20. The jumps can be explained by the fact that to construct the curves we only take triads in which  $A$  sent messages on days  $d$  and  $2d$  and therefore there is an increased likelihood that a fraction of those messages were sent to  $B$ . However, we do not see such jumps on day  $d$  on the red curve even though node  $A$  was active on that day in the  $d$ -closed triads as well as the  $d$ -open triads. The only difference between the red and the green curves is that on day  $d$ ,  $A$  messaged a neighbor of  $B$  in the red curve, but in the green curve  $A$  messaged a node  $D$ , not connected to  $B$ . The lack of jump on day  $d$  can be observed on all the plots of Figure 5.4. This suggests that the communication from  $A$  to  $B$  is in some sense suppressed on the day of the triad's closure, and hence points to a possible case in which  $A$ 's actions toward others are reducing the level of activity on the  $A$ - $B$  link. Understanding the extent of this effect and the mechanism behind it is an intriguing open question.

## 5.4 Basic Properties of Relationship Decay

Since much of our analysis involves the basic fact that interactions on Twitter decay over time, making sustained ties hard to maintain, we now explore the

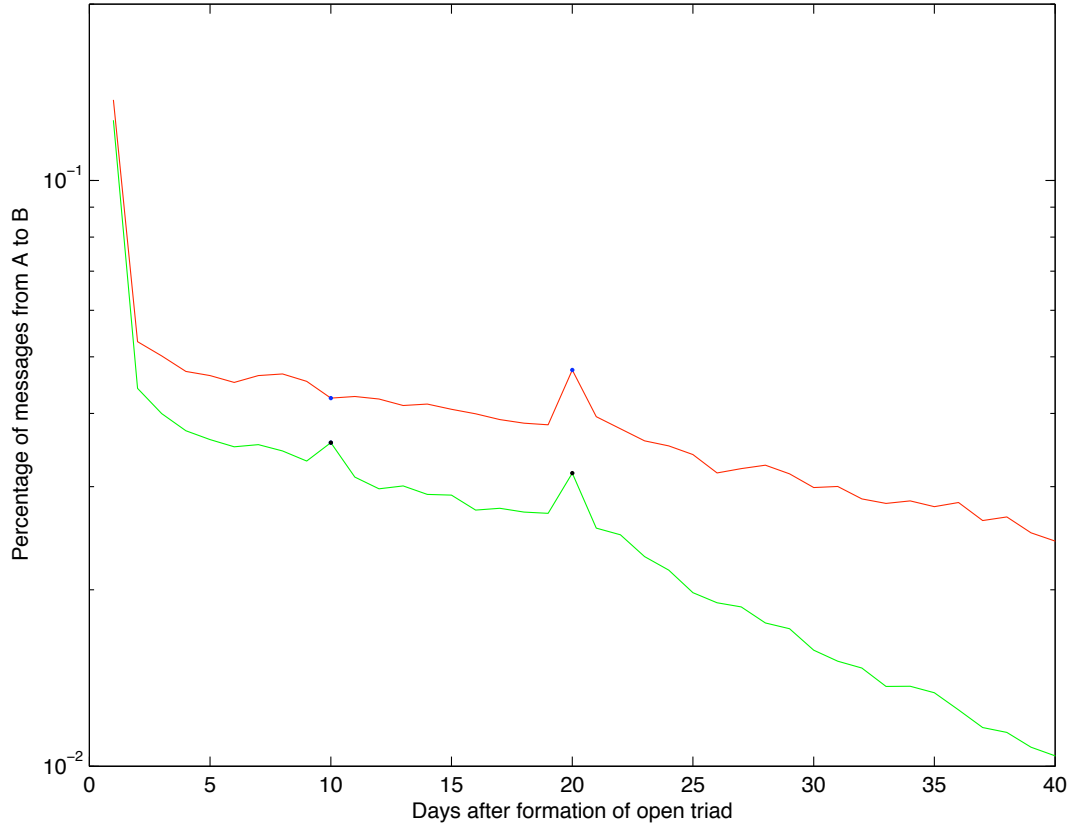


Figure 5.8: Zoom-in of figure 5.4(d). We observe jumps on the green curve at days  $d$  and  $2d$  and on the red curve at day  $2d$  but not on day  $d$ .

basic properties of relationship decay in more detail.

We begin with a simple question. If we observe an event in which  $A$  communicates with  $B$  on day 0, what is the probability that we will observe another such  $A$ -to- $B$  communication event on day  $d > 0$ ? Figure 5.9 shows how this probability decreases as a function of  $d$ : note that it starts with a decay rate that is slower than exponential (though also not a very close fit to a power law), and then straightens out into an approximately exponential rate. We note that the rate of decay is faster than for the curves in Figure 5.4, which were based on

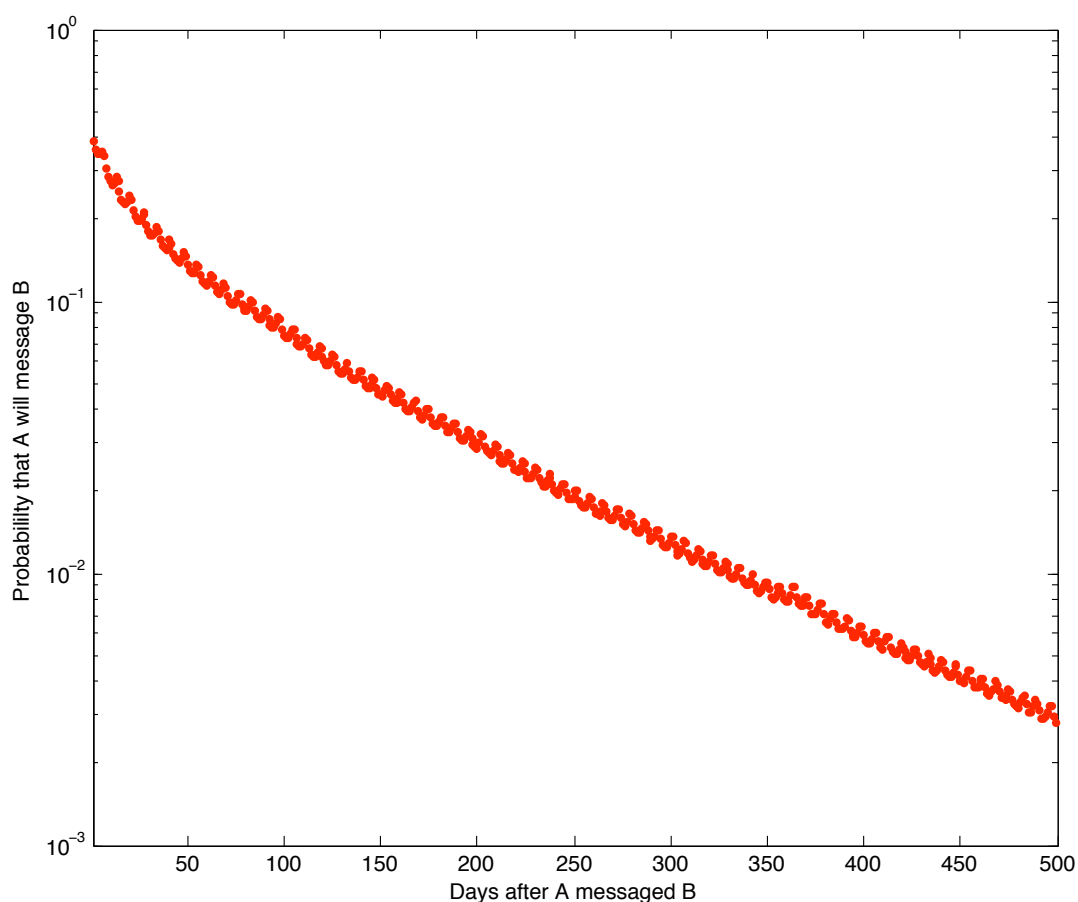


Figure 5.9: Probability that A will send a message to B  $d$  days after having sent her one

$d$ -open and  $d$ -closed triads. Hence, users  $A$  involved in triads tend to maintain their communication with users  $B$  more than the average. One possible interpretation is that their involvement in triads is indicative of a higher level of activity on Twitter overall, triggering the types of spill-over effects that were the focus of the previous section.

The probability of seeing future communication based on just a single observation is one extreme in this genre of questions. At the other extreme, we can study the dynamics of a strong relationship from one user to another, in

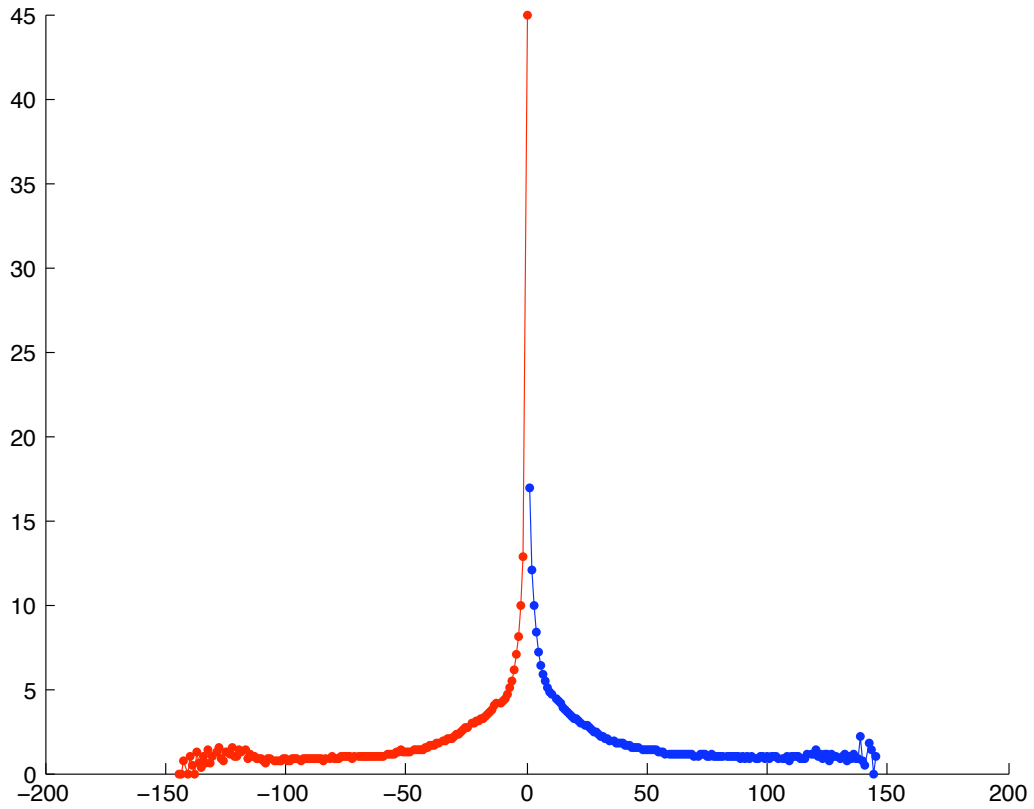


Figure 5.10: Average function  $M$  for pairs  $(A, B)$  in which  $A$  sent  $B$  at least 100 @-messages

which many messages are sent. Specifically, let's consider a pair of users  $(A, B)$  for which  $A$  sends  $B$  at least 100 messages total. We then investigate how the amount of communication from  $A$  to  $B$  changes over time. For each such  $(A, B)$  pair, we partition time into bins of length one week and look for the week during which  $A$  sent the most @-messages to  $B$ . This is the *peak* of the communication. We define the function  $M$  as follows:

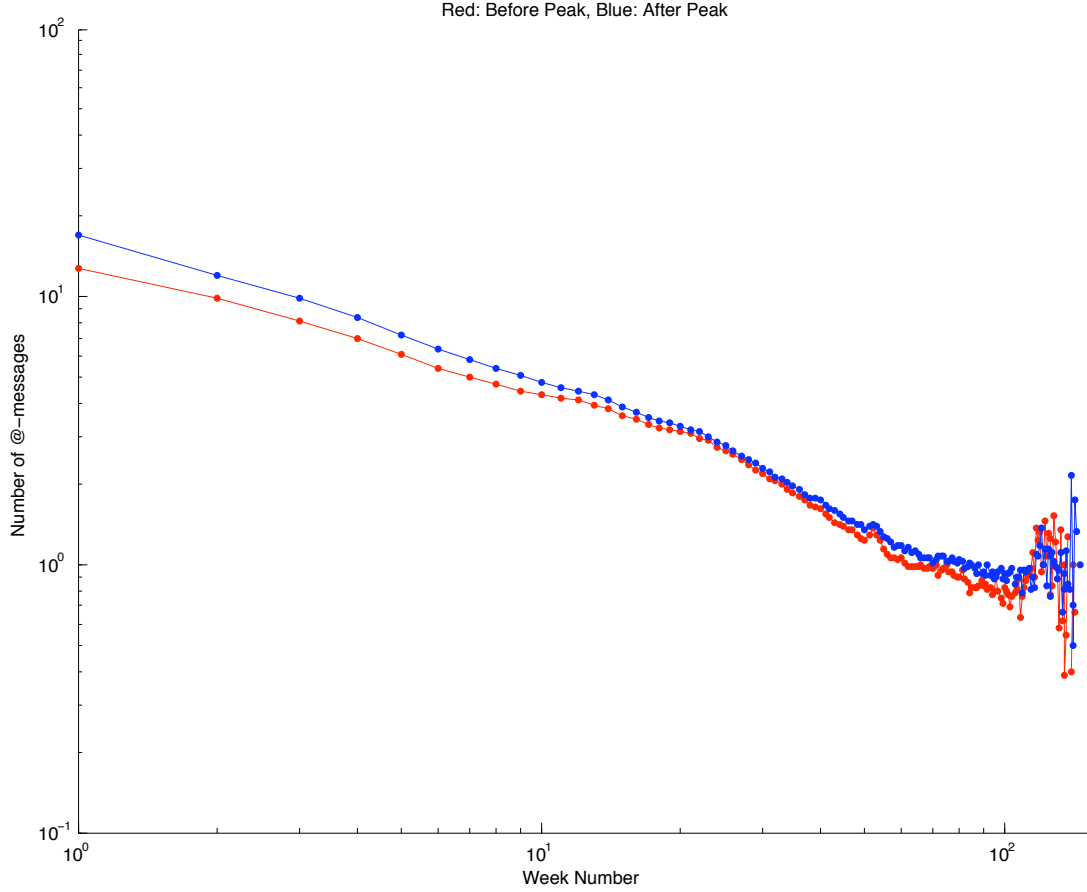


Figure 5.11: Average of  $\log(M(n))$  as a function of  $\log(n)$  where  $n > 0$  (in blue) and as a function of  $\log(-n)$  for  $n < 0$  (in red)

$$M(n) = \begin{cases} \text{Num @-mess during peak} & \text{if } n = 0 \\ \text{Num @-mess during } n^{\text{th}} \text{ week before the peak} & \text{if } n < 0 \\ \text{Num @-mess during } n^{\text{th}} \text{ week after the peak} & \text{if } n > 0 \end{cases}$$

Figure 5.10 shows the average function  $M$  for pairs  $(A, B)$  in which  $A$  sent  $B$  at least 100 @-messages. Figure 5.11 shows the average of  $\log(M(n))$  as a function of  $\log(n)$  where  $n > 0$  (in blue) and as a function of  $\log(-n)$  for  $n < 0$  (in red). The blue curve is above the red curve which suggests that the communication

between  $A$  and  $B$  tends to ramp up to the peak faster than it decays from the peak.

We can refine this analysis a bit further as follows. When a user  $A$  sends a large number of messages to a user  $B$ , there are two possibilities: (i) it could be that  $B$  never send any messages to  $A$  (perhaps because  $B$  is simply a celebrity that  $A$  mentions on a regular basis); or (ii) it could be because  $A$  and  $B$  are actually exchanging messages, suggesting a more overtly social form of interaction. With this in mind, we can consider the plot in Figure 5.10 broken down separately based on whether the messages from  $A$  to  $B$  are *reciprocated* (with  $B$  messaging  $A$  as well) or *unreciprocated* (with no  $B$ -to- $A$  messages). Figure 5.12 shows the results for these two categories: we find that the rates of ramp-up and decay do in fact differ between the two, with the curves for unreciprocated links lying slightly above the corresponding curves for reciprocated links. This suggests that the ramp-up and ramp-down for reciprocated links is in fact slightly more abrupt than it is in the unreciprocated case.

## 5.5 Discussion

There are many forces that affect the strength and longevity of ties on social media sites, and it is a challenge to separate these into their distinct effects. In this chapter we have offered a set of data analysis methodologies that lets us begin to isolate the effect of three such forces: balance, in which ties are strengthened when they close triads; exchange, in which ties are weakened when one end of the tie has other opportunities; and betweenness, in which ties are strengthened when they serve as conduits for information.



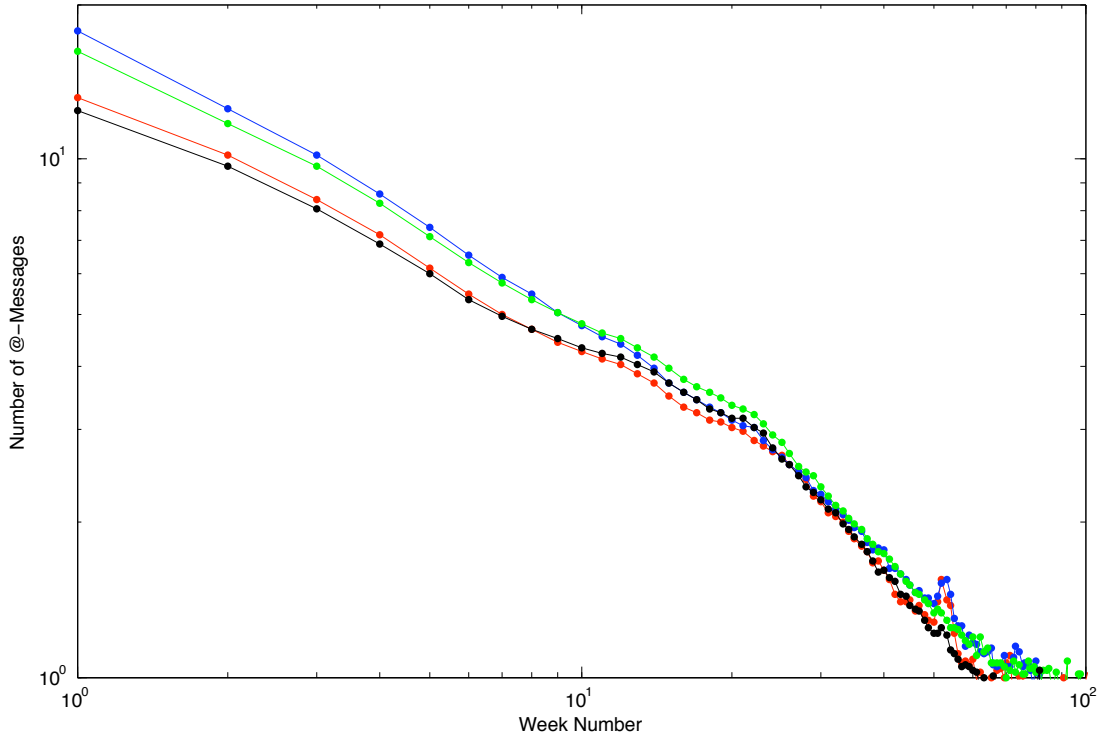


Figure 5.12: Average of  $\log(M(n))$  as a function of  $\log(n)$  where  $n > 0$  (in blue for unreciprocated links and green for reciprocated ones) and as a function of  $\log(-n)$  for  $n < 0$  (in red for unreciprocated links and black for reciprocated ones)

Our analyses show the power of balance in the domain we study, Twitter. It also shows that exchange theory should be broadened to conceptually include off-site opportunities for participants in a tie, reflecting the rapid rate at which ties decay. We believe that the framework developed here can be applied to social media settings quite broadly. In particular, it could be used to analyze the differential rates and trajectories by which relationships grow and decay across different domains, and more intriguingly, it could expose contrasting relative extents to which balance, exchange, and betweenness apply across domains. Ultimately, being able to characterize different social applications through the

different ways in which these forces operate could provide a useful framework for modeling and reasoning about the behavior of these applications.

## **Part II**

# **Information Difussion**

## CHAPTER 6

### THE MECHANICS OF INFORMATION FLOW BY TOPIC

#### 6.1 Introduction

A growing line of recent research has studied the spread of information on-line, investigating the tendency for people to engage in activities such as forwarding messages, linking to articles, joining groups, purchasing products, or becoming fans of pages after some number of their friends have done so [3, 8, 26, 28, 49, 80, 86, 89, 120]. The work in this area has thus far focused primarily on identifying properties that generalize across different domains and different types of information, leading to principles that characterize the process of on-line information diffusion and drawing connections with sociological work on the *diffusion of innovations* [110, 119].

As we begin to understand what is common across different forms of on-line information diffusion, however, it becomes increasingly important to ask about the sources of variation as well. The variations in how different ideas spread is a subject that has attracted the public imagination in recent years, including best-selling books seeking to elucidate the ingredients that make an idea “sticky,” facilitating its spread from one person to another [38, 52]. But despite the fascination with these questions, we do not have a good quantitative picture of how this variation operates at a large scale.

Here are some basic open questions concerning variation in the spread of on-line information. First, the intuitive notion of “stickiness” can be modeled in an idealized form as a probability — the probability that a piece of information will

pass from a person who knows or mentions it to another person who is exposed to it. Are simple differences in the value of this probability indeed the main source of variation in how information spreads? Or are there more fundamental differences in the mechanics of how different pieces of information spread? And if such variations exist at the level of the underlying mechanics, can differences in the type or topic of the information help explain them?

**Variation in the spread of hashtags** In this chapter we analyze sources of variation in how the most widely-used hashtags on Twitter spread within its user population. We find that these sources of variation involve not just differences in the probability with which something spreads from one person to another — the quantitative analogue of stickiness — but also differences in a quantity that can be viewed as a kind of “persistence,” the relative extent to which repeated exposures to a piece of information continue to have significant marginal effects on its adoption.

Moreover, these variations are aligned with the topic of the hashtag. For example, we find that hashtags on politically controversial topics are particularly persistent, with repeated exposures continuing to have large relative effects on adoption; this provides, to our knowledge, the first large-scale validation of the “complex contagion” principle from sociology, which posits that repeated exposures to an idea are particularly crucial when the idea is in some way controversial or contentious [21, 22].

Our data is drawn from a large snapshot of Twitter containing large coverage of all tweets during a period of multiple months. From this dataset, we build a network on the users from the structure of interaction via @-messages; for users

$X$  and  $Y$ , if  $X$  includes “@ $Y$ ” in at least  $t$  tweets, for some threshold  $t$ , we include a directed edge from  $X$  to  $Y$ . @-messages are used on Twitter for a combination of communication and name-invocation (such as mentioning a celebrity via @, even when there is no expectation that they will read the message); under all these modalities, they provide evidence that  $X$  is paying attention to  $Y$ , and with a strength that can be tuned via the parameter  $t$ .<sup>1</sup>

For a given user  $X$ , we call the set of other users to whom  $X$  has an edge the *neighbor set* of  $X$ . As users in  $X$ ’s neighbor set each mention a given hashtag  $H$  in a tweet for the first time, we look at the probability that  $X$  will first mention it as well; in effect, we are asking, “How do successive exposures to  $H$  affect the probability that  $X$  will begin mentioning it?” Concretely, following the methodology of [26], we look at all users  $X$  who have not yet mentioned  $H$ , but for whom  $k$  neighbors have; we define  $p(k)$  to be the fraction of such users who mention  $H$  before a  $(k + 1)^{\text{st}}$  neighbor does so. In other words,  $p(k)$  is the fraction of users who adopt the hashtag directly after their  $k^{\text{th}}$  “exposure” to it, given that they hadn’t yet adopted it.

As an example, Figure 6.1 shows a plot of  $p(k)$  as a function of  $k$  averaged over the 500 most-mentioned hashtags in our dataset. Note that these top hashtags are used in sufficient volume that one can also construct meaningful  $p(k)$  curves for each of them separately, a fact that will be important for our subsequent analysis. For now, however, we can already observe two basic features of the average  $p(k)$  curve’s shape: a ramp-up to a peak value that is reached rela-

---

<sup>1</sup>One can also construct a directed network from the *follower* relationship, including an edge from  $X$  to  $Y$  if  $X$  follows  $Y$ . We focus here on @-messages in part because of a data resolution issues — they can be recovered with exact time stamps from the tweets themselves — but also because of earlier research suggesting that users often follow other users in huge numbers and hence potentially less discriminately, whereas interaction via @-messages indicates a kind of attention that is allocated more parsimoniously, and with a strength that can be measured by the number of repeat occurrences [59].

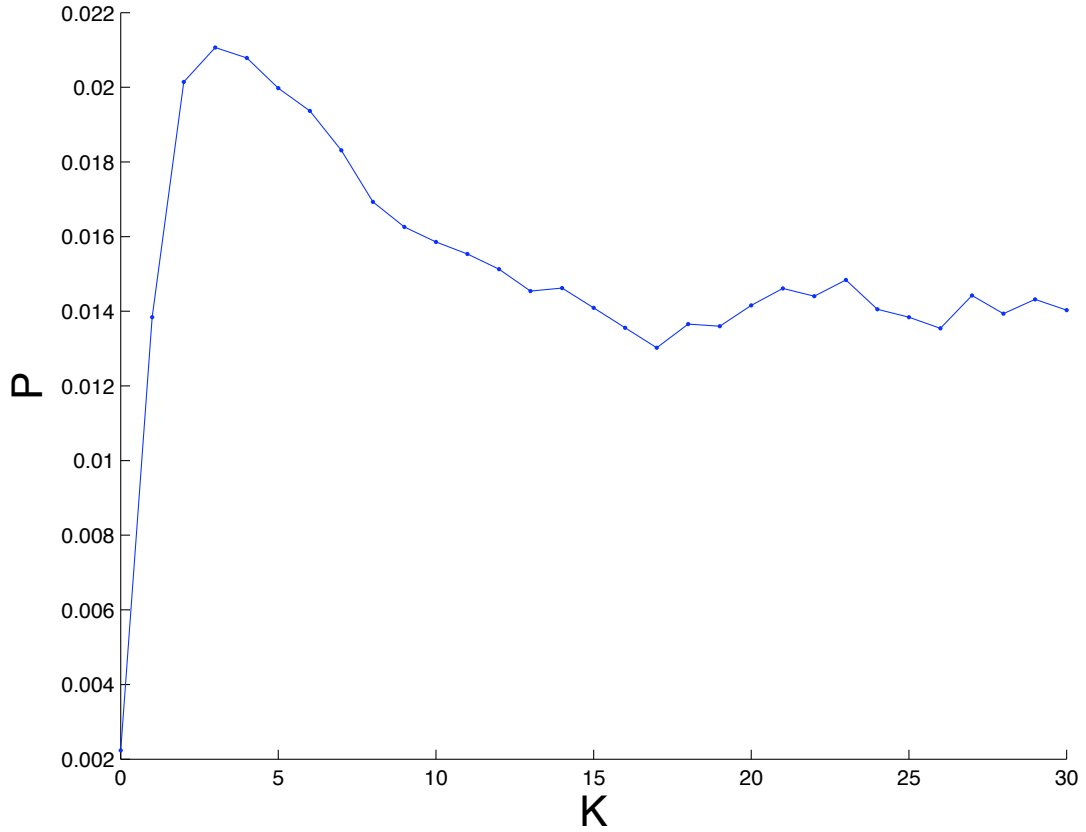


Figure 6.1: Average exposure curve for the top 500 hashtags.  $P(K)$  is the fraction of users who adopt the hashtag directly after their  $k^{\text{th}}$  exposure to it, given that they had not yet adopted it

tively early (at  $k = 2, 3, 4$ ), followed by a decline for larger values of  $k$ . In keeping with the informal discussion above, we define the *stickiness* of the curve to be the maximum value of  $p(k)$  (since this is the maximum probability with which an exposure to  $H$  transfers to another user), and the *persistence* of the curve to be a measure of its rate of decay after the peak.<sup>2</sup> We will find that, in a precise sense, these two quantities — stickiness and persistence — are sufficient to approximately characterize the shapes of individual  $p(k)$  curves.

<sup>2</sup>We formally define persistence in Section 6.3; roughly, it is the ratio of the area under the curve to the area of the largest rectangle that can be circumscribed around it.

**Variation in Adoption Dynamics Across Topics** The shape of  $p(k)$  averaged over all hashtags is similar to analogous curves measured recently in other domains [26], and our interest here is in going beyond this aggregate shape and understanding how these curves vary across different kinds of hashtags. To do this, we first classified the 500 most-mentioned hashtags according to their topic. We then average the curves  $p(k)$  separately within each category and compare their shapes.<sup>3</sup>

Many of the categories have  $p(k)$  curves that do not differ significantly in shape from the average, but we find unusual shapes for several important categories. First, for political hashtags, the persistence has a significantly larger value than the average — in other words, successive exposures to a political hashtag have an unusually large effect relative to the peak. This is striking in the way that it accords with the “complex contagion” principle discussed earlier: when a particular behavior is controversial or contentious, people may need more exposure to it from others before adopting it themselves [21, 22].

In contrast, we find a different form of unusual behavior from a class of hashtags that we refer to as Twitter *idioms* — a kind of hashtag that will be familiar to Twitter users in which common English words are concatenated together to serve as a marker for a conversational theme (e.g. #cantlivewithout, #dontyouhate, #iloveitwhen, and many others, including concatenated markers for weekly Twitter events such as #musicmonday and #followfriday.) Here the

---

<sup>3</sup>In Section 6.2 we describe the methodology used to perform this manual classification in detail. In brief, we compared independent classifications of the hashtags obtained by disjoint means, involving annotation by the authors compared with independent annotation by a group of volunteers. Our results based on the average curves arising from this classification are robust in the following sense: despite differences in classification of some individual hashtags by the two groups, the curves themselves exhibit essentially identical behavior when computed from either of the two classifications separately, as well as from an intersection of the two classifications.



stickiness is high, but the persistence is unusually low; if a user doesn't adopt an idiom after a small number of exposures, the marginal chance they do so later falls off quickly.

**Subgraph Structure and Tie Strength** In addition to the person-to-person mechanics of spread, it is also interesting to look at the overall structure of interconnections among the initial adopters of a hashtag. To do this, we take the first  $m$  individuals to mention a particular hashtag  $H$ , and we study the structure of the subgraph  $G_m$  induced on these first  $m$  mentioners. In this structural context, we again find that political hashtags exhibit distinctive features — in particular, the subgraphs  $G_m$  for political hashtags  $H$  tend to exhibit higher internal degree, a greater density of triangles, and a large number of nodes not in  $G_m$  who have significant numbers of neighbors in it. This is again broadly consistent with the sociological premises of complex contagion, which argues that the successful spread of controversial behaviors requires a network structure with significant connectivity and significant local clustering.

Within these subgraphs, we can consider a set of sociological principles that are related to complex contagion but distinct from it, centered on the issue of *tie strength*. Work of McAdam and others has argued that the sets of early adopters of controversial or risky behaviors tend to be rich in strong ties, and that strong ties are crucial for these activities [95, 96] — in contrast to the ways in which learning about novel information can correspondingly benefit from transmission across weaker ties [44].

When we look at tie strength in these subgraphs, we find a somewhat complex picture. Because subgraphs  $G_m$  for political hashtags have significantly

more edges, they have more ties of all strengths, including strong ties (according to several different definitions of strength summarized in Section 6.4). This aspect of the data aligns with the theories of McAdam and others. However, the fraction of strong ties in political subgraphs  $G_m$  is actually *lower* than the fraction of strong ties for the full population of widely-used hashtags, indicating the overall greater density of edges in political subgraphs comes more dominantly from a growth in weak ties than from strong ones. The picture that emerges of early-adopter subgraphs for political hashtags is thus a subtle one: they are structures whose communication patterns are more densely connected than the early-adopter subgraphs for other hashtags, and this connectivity comes from a core of strong ties embedded in an even larger profusion of weak ties.

**Interpreting the Findings** When we look at politically controversial topics on Twitter, we therefore see both direct reflections and unexpected variations on the sociological theories concerning how such topics spread. This is part of a broader and important issue: understanding differences in the dynamics of contentious behavior in the off-line world versus the on-line world. It goes without saying that the use of a hashtag on Twitter isn't in any sense comparable, in terms of commitment or personal risk, to taking part in activism in the physical world (a point recently stressed in a much-circulated article by Malcolm Gladwell [39]). But the underlying issue persists on Twitter: political hashtags are still riskier to use than conversational idioms, albeit at these much lower stakes, since they involve publicly aligning yourself with a position that might alienate you from others in your social circle. The fact that we see fundamental aspects of the same sociological principles at work both on-line and off-line suggests a certain robustness to these principles, and the differences that we see suggest a

perspective for developing deeper insights into the relationship between these behaviors in the on-line and off-line domains.

This distinction between contentious topics in the on-line and off-line worlds is one issue to keep in mind when interpreting these results. Another is the cumulative nature of the findings. As with any analysis at this scale, we are not focusing on why any one individual made the decisions they did, nor is it the case that that Twitter users are even aware of all the tweets containing their exposures to hashtags via neighbors. Rather, the point is that we still find a strong signal in an aggregate sense — as a whole, the population is exhibiting differences in how it responds to hashtags of different types, and in ways that accord with theoretical work in other domains.

A further point to emphasize is that our focus in this work is on the hashtags that succeeded in reaching large numbers of people. It is an interesting question to consider what distinguishes a hashtag that spreads widely from one that fails to attract attention, but that is not the central question we consider here. Rather, what we are identifying is that among hashtags that do reach many people, there can nevertheless be quite different mechanisms of contagion at work, based on variations in stickiness and persistence, and that these variations align in interesting ways with the topic of the hashtag itself.

**Simulated Spreading** Finally, an interesting issue here is the interaction between the  $p(k)$  curve and the subgraph  $G_m$  for a given hashtag  $H$  — clearly the two develop in a form of co-evolution, since the addition of members via the curve  $p(k)$  determines how the subgraph of adopters takes shape, but the structure of this subgraph — particularly in the connections between adopters and

non-adopters — affects who is likely to use the hashtag next. To understand how  $p(k)$  and  $G_m$  relate to each other, it is natural to consider questions of the following form: how would the evolution of  $G_m$  have turned out differently if a different  $p(k)$  curve had been in effect? Or correspondingly, how effectively would a hashtag with curve  $p(k)$  have spread if it had started from a different subgraph  $G_m$ ? Clearly it is difficult to directly perform this counterfactual experiment as stated, but we obtain insight into the structure of the question by simulating the  $p(k)$  curve of each top hashtag on the subgraph  $G_m$  of each other top hashtag. In this way, we begin to identify some of the structural factors at work in the interplay between the mechanics of person-to-person influence and the network on which it is spreading.

## 6.2 Dataset, Network Definition, and Hashtag Classification

**Data Collection and Network Definition** We used the data set described in chapter 5. As discussed earlier, in addition to extracting tweets and hashtags within them, we also build a network on the users, connecting user  $X$  to user  $Y$  if  $X$  directed at least  $t$  @-messages to  $Y$ . In our analyses we use  $t = 3$ , except when we are explicitly varying this parameter. There are multiple ways of defining a network on which hashtags can be viewed as diffusing, and our definition is one way of defining a proxy for the attention that users  $X$  pay to other users  $Y$ .

**Hashtag Selection and Classification** To create a classification of hashtags by category, we began with the 500 hashtags in the data that had been mentioned by the most users. From manual inspection of this list, we identified eight broad

Category	Definition
Celebrity	The name of a person or group (e.g. music group) that is featured prominently in entertainment news. Political figures or commentators with a primarily political focus are not included. The name of the celebrity may be embedded in a longer hashtag referring to some event or fan group that involves the celebrity. Note that many music groups have unusual names; these still count under the “celebrity” category.
Games	Names of computer, video, MMORPG, or twitter-based games, as well as groups devoted to such games.
Idiom	A tag representing a conversational theme on twitter, consisting of a concatenation of at least two common words. The concatenation can’t include names of people or places, and the full phrase can’t be a proper noun in itself (e.g. a title of a song/movie/organization). Names of days are allowed in the concatenation, because of the the Twitter convention of forming hashtags involving names of days (e.g. MusicMonday). Abbreviations are allowed only if the full form also appears as a top hashtag (so this rules out hashtags including omg, wtf, lol, nsfw).
Movies/TV	Names of movies or TV shows, movie or TV studios, events involving a particular movie or TV show, or names of performers who have a movie or TV show specifically based around them. Names of people who have simply appeared on TV or in a movie do not count.
Music	Names of songs, albums, groups, movies or TV shows based around music, technology designed for playing music, or events involving any of these. Note that many music groups have unusual names; these still count under the “music” category.
Political	A hashtag that in your opinion often refers to a politically controversial topic. This can include a political figure, a political commentator, a political party or movement, a group on twitter devoted to discussing a political cause, a location in the world that is the subject of controversial political discussion, or a topic or issue that is the subject of controversial political discussion. Note that this can include political hashtags oriented around countries other than the U.S.
Sports	Names of sports teams, leagues, athletes, particular sports or sporting events, fan groups devoted to sports, or references to news items specifically involving sports.
Technology	Names of Web sites, applications, devices, or events specifically involving any of these.

Table 6.1: Definitions of categories used for annotation.

Category	Examples	Category	Examples
Celebrity	mj, brazilwantsjb, regis, iwantpeter-facinelli	Music	thisiswar, mj, musicmonday, pandora
Games	mafiawars, spymaster, mw2, zyn-gapirates	Political	tcot, glennbeck, obama, hcr
Idiom	cantlivewithout, dontyouhate, musicmonday	Sports	golf, yankees, nhl, cricket
Movies/TV	lost, glennbeck, bones, newmoon	Technology	digg, iphone, jquery, photoshop

Table 6.2: A small set of examples of members in each category.

categories of hashtags that each had at least 20 clear exemplars among these top hashtags, and in most cases significantly more. (Of course, many of the top 500 hashtags fit into none of the categories.) We formulated definitions of these categories as shown in Table 6.1. Then we applied multiple independent mechanisms for classifying the hashtags according to these categories. First, the

authors independently annotated each hashtag, and then had a reconciliation phase in which they noted errors and arrived at a majority judgment on each annotation. Second, the authors solicited a group of independent annotators, and took the majority among their judgments. Annotators were provided with the category definitions, and for each hashtag were provided with the tag's definitions (when present) from the Web resources Wthashtag and Tagalus, as well as links to Google and Twitter search results on the tag. Finally, since the definition of the "idiom" category is purely syntactic, we did not use annotators for this task, but only for the other seven categories.

Clearly even with this level of specificity, involving both human annotation and Web-based definitional resources, there are ultimately subjective judgments involved in category assignments. However, given the goal of understanding variations in hashtag behavior across topical categories, at some point in the process a set of judgments of this form is unavoidable. What we find is the results are robust in the presence of these judgments: the level of agreement among annotators was uniformly high, and the plots presented in the subsequent sections show essentially identical behavior regardless of whether they are based on the authors' annotations, the independent volunteers' annotations, or the intersection of the two. To provide the reader with some intuition for the kinds of hashtags that fit each category, we present a handful of illustrative examples in Table 6.2, drawn from the much larger full membership in each category. The full category memberships can be seen at <http://www.cam.cornell.edu/~dromero/top500ht>.

## 6.3 Exposure Curves

**Basic definitions** In order to investigate the mechanisms by which hashtag usage spreads among Twitter users, we begin by reviewing two ways of measuring the impact that exposure to others has in an individual's choice to adopt a new behavior (in this case, using a hashtag) [26]. We say that a user is *k-exposed* to hashtag *h* if he has not used *h*, but has edges to *k* other users who have used *h* in the past. Given a user *u* that is *k-exposed* to *h* we would like to estimate the probability that *u* will use *h* in the future. Here are two basic ways of doing this.

**Ordinal time estimate.** Assume that user *u* is *k-exposed* to some hashtag *h*. We will estimate the probability that *u* will use *h* before becoming  $(k + 1)$ -exposed. Let  $E(k)$  be the number of users who were *k-exposed* to *h* at some time, and let  $I(k)$  be the number of users that were *k-exposed* and used *h* before becoming  $(k + 1)$ -exposed. We then conclude that the probability of using the hashtag *h* while being *k-exposed* to *h* is  $p(k) = \frac{I(k)}{E(k)}$ .

**Snapshot estimate.** Given a time interval  $T = (t_1, t_2)$ , assume that a user *u* is *k-exposed* to some hashtag *h* at time  $t = t_1$ . We will estimate the probability that *u* will use *h* sometime during time interval *T*. We let  $E(k)$  be the number of users who were *k-exposed* to *h* at time  $t = t_1$ , and let  $I(k)$  be the number of users who were *k-exposed* to *h* at time  $t = t_1$  and used *h* sometime before  $t = t_2$ . We then conclude that  $p(k) = \frac{I(k)}{E(k)}$  is the probability of using *h* before time  $t = t_2$ , conditioned on being *k-exposed* to *h* at time  $t = t_1$ . We will refer to  $p(k)$  as an *exposure curve*; we will also informally refer to it as an *influence curve*, although it is being used only for prediction, not necessarily to infer causal influence.

The ordinal time approach requires more detailed data than the snapshot

method. Since our data are detailed enough that we are able to generate the ordinal time estimate, we only present the results based on the ordinal time approach; however, we have confirmed that the conclusions hold regardless of which approach is followed. This is not surprising since it has been argued that sufficiently many snapshot estimates contain enough information to infer the ordinal time estimate [26].

**Comparison of Hashtag Categories: Persistence and Stickiness** We calculated ordinal time estimates  $P(k)$  for each one of the 500 hashtags we consider. For each point on each curve we calculate the 95% Binomial proportion confidence interval. We observed some qualitative differences between the curves corresponding to different hashtags. In particular, we noticed that some curves increased dramatically initially as  $k$  increased but then started to decrease relatively fast, while other curves increased at a much slower rate initially but then saturated or decreased at a much slower rate. As an example, Figure 6.3 shows the influence curves for the hashtags #cantlivewithout and #hcr. We also noticed that some curves had much higher maximum values than others.<sup>4</sup>

In this discussion, we are basing differences among hashtags on different structural properties of their influence curves. In order to make these distinctions more precise we use the following measures.

First, we formalize a notion of “persistence” for an influence curve, capturing how rapidly it decays. Formally, given a function  $P : [0, K] \rightarrow [0, 1]$  we let  $R(P) = K \max_{k \in [0, K]} \{P(k)\}$  be the area of the rectangle with length  $K$  and height

---

<sup>4</sup>As  $k$  gets larger the amount of data used to calculate  $P(k)$  decreases, making the error intervals very large and the curve very noisy. In order to take this into account we only defined  $P(k)$  when the relative error was less than some value  $\theta$ . Throughout the study we checked that the results held for different values of  $\theta$ .



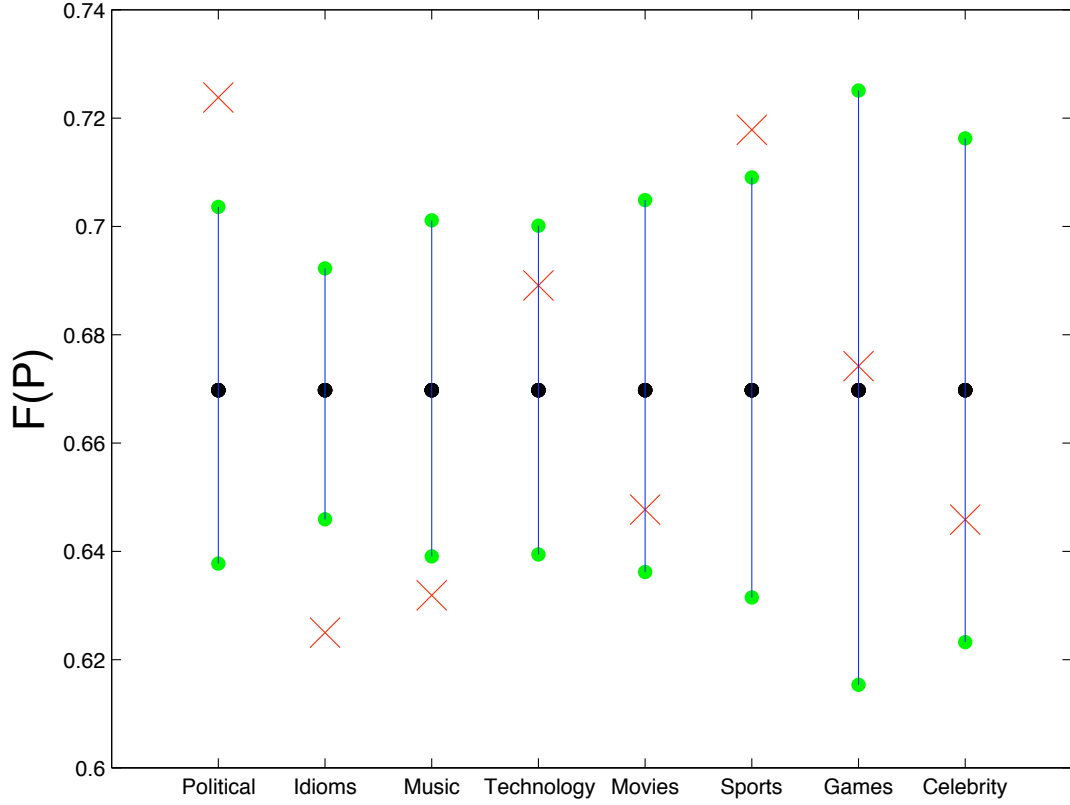


Figure 6.2:  $F(P)$  for the different types of hashtags. The black dots are the average  $F(P)$  among all hashtags, the red x is the average for the specific category, and the green dots indicate the 90% expected interval where the average for the specific set of hashtags would be if the set was chosen at random. Each point is the average of a set of at least 10 hashtags

$\max_{k \in [0, K]} \{P(k)\}$ . We let  $A(P)$  be the area under the curve  $P$  assuming the point  $P(k)$  is connected to the point  $P(k + 1)$  by a straight line. Finally, we let  $F(P) = \frac{A(P)}{R(P)}$  be the *persistence* parameter.

When an influence curve  $P$  initially increases rapidly and then decreases, it will have a smaller value of  $F(P)$  than a curve  $\tilde{P}$  which increases slowly and then saturates. Similarly, an influence curve  $P$  that slowly increases monoton-

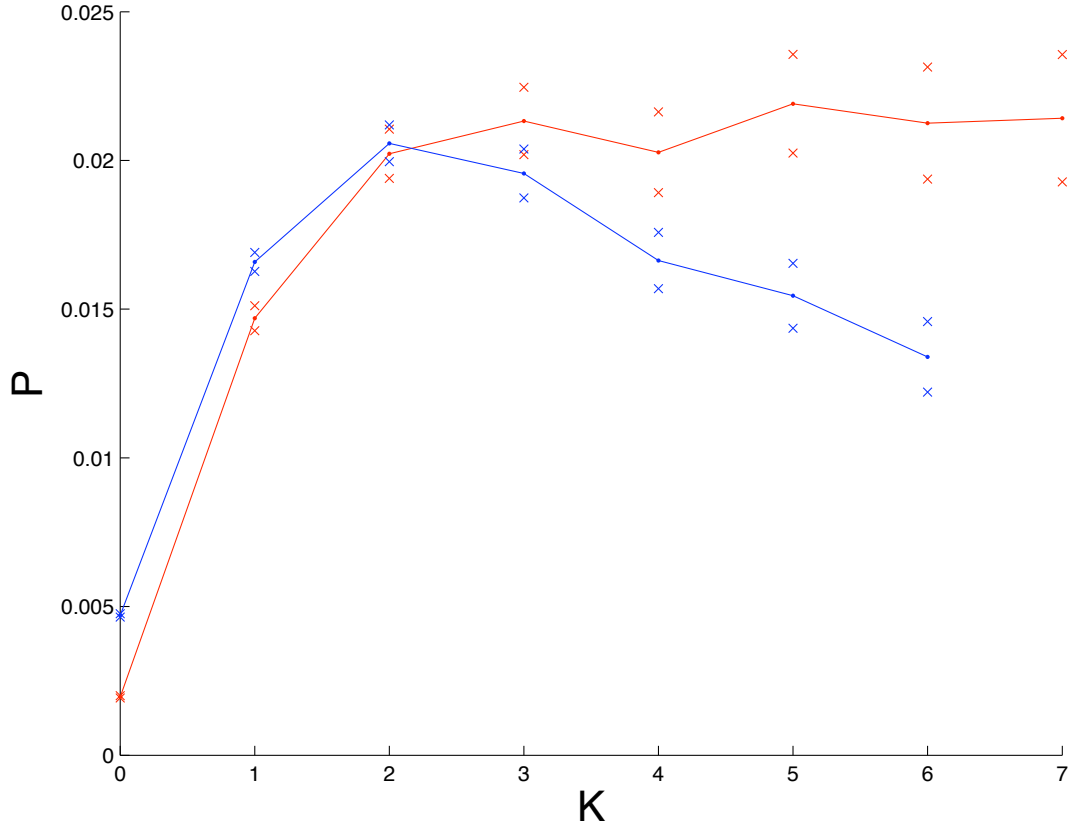


Figure 6.3: Sample exposure curves for hashtags #cantlivewithout (blue) and #hcr (red).

ically will have a smaller value of  $F(P)$  than a curve  $\tilde{P}$  that initially increases rapidly and then saturates. Hence the measure  $F$  captures some differences in the shapes of the influence curves. In particular, applying this measure to an influence curve would tell us something about its persistence; the higher the value of  $F(P)$ , the more persistent  $P$  is.

Second, given an influence curve  $P : [0, K] \rightarrow [0, 1]$  we let  $M(P) = \max_{k \in [0, K]} \{P(k)\}$  be the *stickiness* parameter, which gives us a sense for how large the probability of usage can be for a particular hashtag based on the most effective exposure.

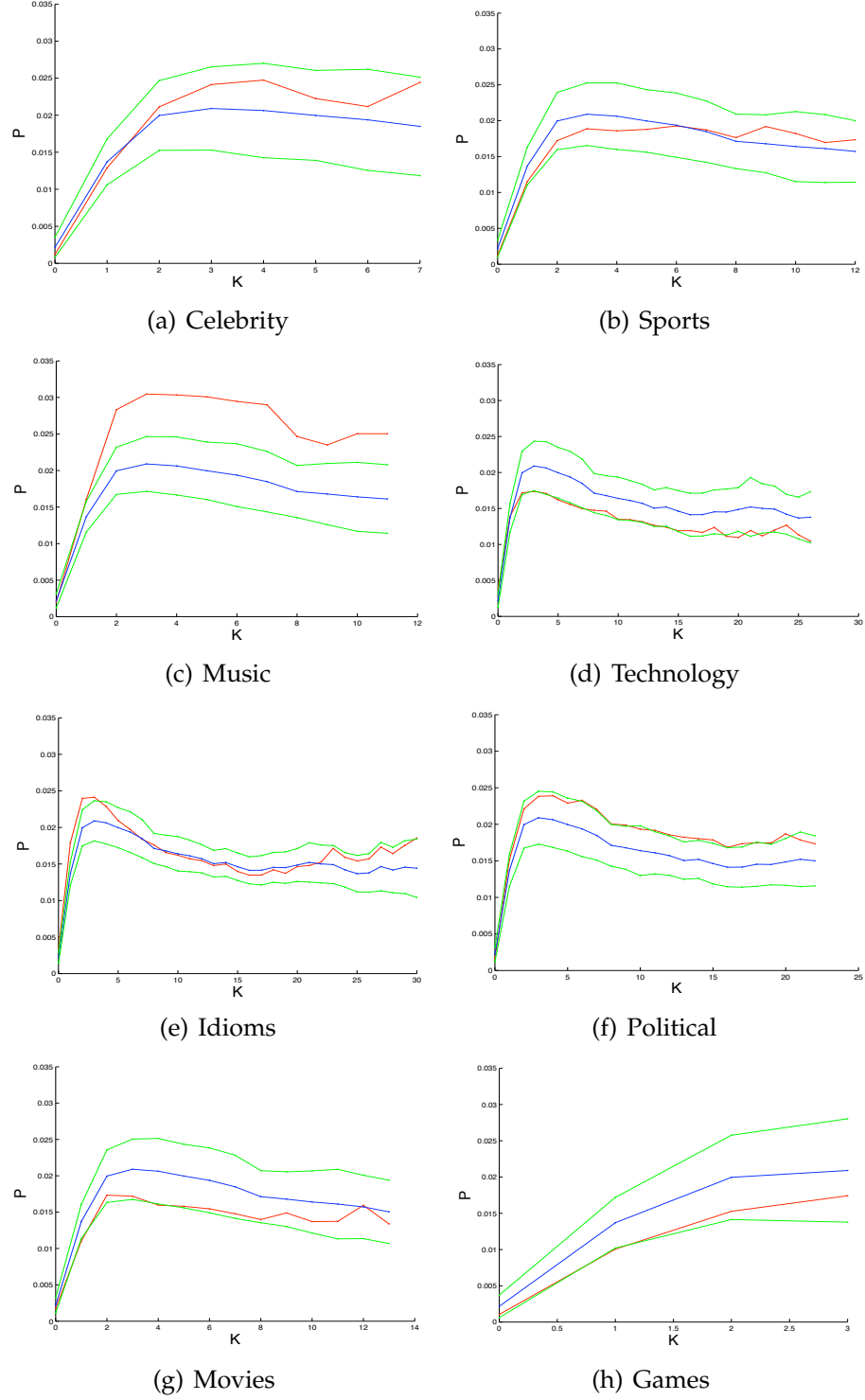


Figure 6.4: Point-wise average influence curves. The blue line is the average of all the influence curves, the red line is the average for the set of hashtags of the particular topic, and the green lines indicate the interval where the red line is expected to be if the hashtags were chosen at random.

We are interested in finding differences between the spreading mechanism of different topics on Twitter. We start by finding out if hashtags corresponding to different topics have influence curves with different shapes. We found significant differences in the values of  $F(P)$  for different topics. Figure 6.2 shows the average  $F(P)$  for the different categories, compared to a baseline in which we draw a set of categories of the same size uniformly at random from the full collection of 500. We see that politics and sports have an average value of  $F(P)$  which is significantly higher than expected by chance, while for Idioms and Music it is lower. This suggests that the mechanism that controls the spread of hashtags related to sports or politics tends to be more persistent than average; repeated exposures to users who use these hashtags affects the probability that a person will eventually use the hashtag more positively than average. On the other hand, for Idioms and Music, the effect of repeated exposures falls off more quickly, relative to the peak, compared to average.

Figure 6.4 shows the point-wise average of the influence curves for each one of the categories. Here we can see some of the differences in persistence and stickiness the curves have. For example, the stickiness of the topics Music, Celebrity, Idioms, and politics tends to be higher than average since the average influence curve for those categories tends to be higher than the average influence curve for all hashtags, while that of Technology, Movies, and Sports tends to be lower than average. On the other hand, these plots give us more intuition on why we found that politics and Sports have a high persistence while for Idioms and Music it is low. In the case of Politics, we see that the red curve starts off just below the green curve (the upper error bar) and as  $k$  increases, the red curve increases enough to be above the green. Similarly, the red curve for Sports starts below the blue curve and it ends above it. In the case of Idioms, the red

curve initially increases rapidly but then it drops below the blue curve. Similarly, the red curve for Music is always very high and above all the other curves, but it drops faster than the other curves at the end.

**Approximating Curves via Stickiness and Persistence** When we compare curves based on their stickiness and persistence, it is important to ask whether these are indeed an adequate pair of parameters for discussing the curves' overall "shapes." We now establish that they are, in the following sense: we show that these two parameters capture enough information about the influence curves that we can approximate the curves reasonably well given just these two parameters. Assume that for some curve  $P$  we are given  $F(P)$  and  $M(P)$ . We will also assume that we know the maximum value of  $k = K$  for which  $P(k)$  is defined. Then we will construct an approximation curve  $\tilde{P}$  in the following way:

1. Let  $\tilde{P}(0) = 0$
2. Let  $\tilde{P}(2) = M(P)$
3. Now we will let  $\tilde{P}(K)$  be such that  $F(\tilde{P}) = F(P)$ . This value turns out to be 
$$\tilde{P}(K) = \frac{M(P) * K * (2 * F(P) - 1)}{K - 2}$$
4. Finally, we will make  $\tilde{P}$  be piecewise linear with one line connecting the points  $(0, 0)$  and  $(2, M(P))$ , and another line connecting the points  $(2, M(P))$  and  $(K, \frac{M(P) * K * (2 * F(P) - 1)}{K - 2})$ .

Figure 6.5 shows an example of an approximation for a particular influence curve. In order to test the quality of the approximation  $\tilde{P}$  we define the approx-

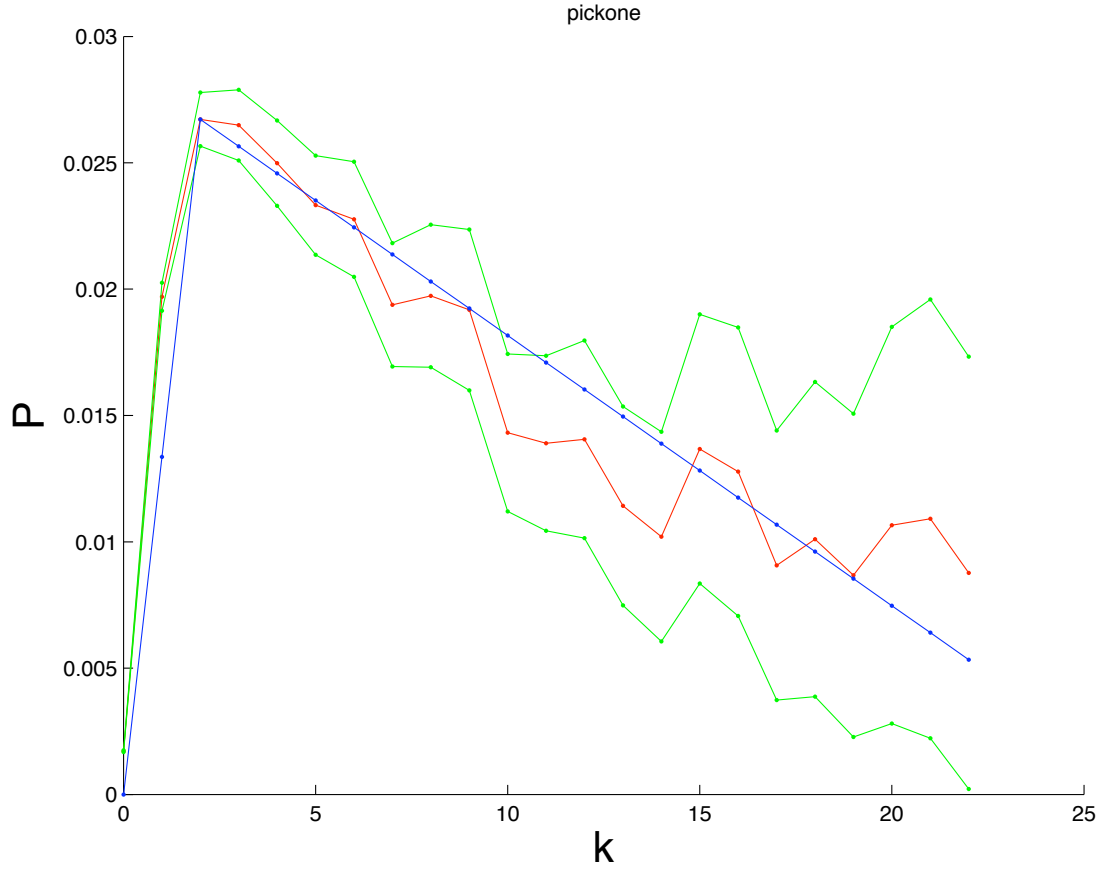


Figure 6.5: Example of the approximation of an influence curve. The red curve is the influence curve for the hashtag #pickone, the green curves indicate the 95% binomial confidence interval, and the blue curve is the approximation.

imation error between  $\tilde{P}$  and  $P$  as the mean absolute error

$$E(P, \tilde{P}) = \frac{1}{K} \sum_{k=0}^K |P(k) - \tilde{P}(k)|$$

and compare it with the mean absolute of the error  $E(P)$  obtained from the 95% confidence intervals around each point  $P(k)$ . The average approximation error among all the influence curves is 0.0056 and the average error of based on the confidence intervals is 0.0050. The approximation error is slightly smaller, which means that out approximation is, on average, within the 95% confidence

interval from the actual influence curve. This suggests the information contained in the stickiness and persistence parameters are enough to accurately approximate the influence curves and gives more meaning to the approach of comparing the curves by comparing these two parameters.

**Frequency of Hashtag Usage** We have observed that different topics have differences in their spreading mechanisms. We also found that they differed in other ways. For example, we see some variation in the number of mentions and the number of users of each category. Table 6.3 shows the different median values for number of mentions, number of users, and number of mentions per user for different types of hashtags. We see that while Idioms and Technology hashtags are used by many users compared to others, each user only uses the hashtag a few times and hence the total number of mentions of the these categories is not much higher than others. On the other hand, only relatively few people used Political and Games hashtags, but each one of them used them many times, making them the most mentioned categories. In the case of games, a contributing factor is that some of users of game hashtags allow external websites to post on their Twitter account every time they accomplish something in the game, which tends to happen very often. It is not clear that there is a correspondingly simple explanation for the large number of mentions per user for political hashtags, but one can certainly conjecture that it may reflect something about the intensity with which these topics are discussed by the users who engage in such discussions; this is an interesting issue to explore further.

Type	Mdn. Mentions	Mdn. Users	Mdn. Ment./User
All HTS	93,056	15,418	6.59
Political	132,180	13,739	10.17
Sports	98,234	11,329	9.97
Idioms	99,317	26,319	3.54
Movies	90,425	15,957	6.57
Celebrity	87,653	5,351	17.68
Technology	90,462	24,648	5.08
Games	123,508	15,325	6.61
Music	87,985	7,976	10.39

Table 6.3: Median values for number of mentions, number of users, and number of mentions per user for different types of hashtags

## 6.4 The structure of initial sets

The spread of a given piece of information is affected by the diffusion mechanism controlled by the influence curves discussed in the previous section, but it may also be affected by the structure of the network relative to the users of the hashtag. To explore this further, we looked at the subgraph  $G_m$  induced by the first  $m$  people who used a given hashtag. We found that there are important differences in the structure of those graphs.

In particular, we consider differences in the structures of the subgraphs  $G_m$  across different categories. For each graph  $G_m$ , across all hashtags and a sequence of values of  $m$ , we compute several structural parameters. First, we compute the average degree of the nodes and the number of triangles in the graph. Then, we defined the *border* of  $G_m$  to be the set of all nodes not in  $G_m$  who have at least one edge to a node in  $G_m$ , and we define the *entering degree* of a node in the border to be the number of neighbors it has in  $G_m$ . We consider the size of the border and the average entering degree of nodes in the border.



Looking across all categories, we find that political hashtags are the category in which the most significant structural differences from the average occur. Table 6.4 shows the averages for political hashtags compared to the average for all hashtags, using the subgraphs  $G_{500}$  on the first 500 users.<sup>5</sup> In brief, the early adopters of a political hashtag message with more people, creating more triangles, and with a border of people who have more links on average into the early adopter set. The number of triangles, in fact, is high even given the high average degree; clearly one should expect a larger number of triangles in a subgraph of larger average degree, but in fact the triangle count for political hashtags is high even when compared against a baseline consisting of non-political hashtags with comparable average degrees. These large numbers of edges and triangles are consistent with the predictions of complex contagion, which argues that such structural properties are important for the spread of controversial topics [22].

**Tie Strength** There is an interesting further aspect to these structural results, obtained by looking at the *strength* of the ties within these subgraphs. There are multiple ways of defining tie strength from social media data [36], and here we consider two distinct approaches. One approach is to use the total number of @-messages sent across the link as a numerical measure of strength. Alternately, we can declare a link to be strong if and only if it is *reciprocated* (i.e. declaring  $(X, Y)$  to be strong if and only if  $(Y, X)$  is in the subgraph as well, following a standard working notion of reciprocation as a proxy for tie strength in the sociology literature [47]).

Under both definitions, we find that the fraction of strong ties in subgraphs

---

<sup>5</sup>The results are similar for  $G_m$  with a range of other values of  $m \neq 500$ .

Type	I	II	III	IV
All HTS	1.41	384	1.24	13425
Political	2.55	935	1.41	12879
Upper Error Bar	1.82	653	1.32	15838
Lower Error Bar	1.00	112	1.16	11016

Table 6.4: Comparison of graphs induced by the first 500 early adopters of political hashtags and average hashtags. Column definitions: I. Average degree, II. Average triangle count, III. Average entering degree of the nodes in the border of the graphs, IV. Average number of nodes in the border of the graphs. The error bars indicate the 95% confidence interval of the average value of a randomly selected set of hashtags of the same size as Political.

$G_m$  for political hashtags is in fact significantly lower than the fraction of strong ties in subgraphs  $G_m$  for our set of hashtags overall. However, since political subgraphs  $G_m$  contain so many links relative to the typical  $G_m$ , we find that they have a larger absolute number of strong ties. As noted in the introduction, standard sociological theories suggest that we should see many strong ties in subgraphs  $G_m$  for political topics, but the picture we obtain is more subtle in that the growth in strong ties comes with an even more significant growth in weak ties. Understanding these competing forces in the structural behavior of such subgraphs is an interesting open question.

## 6.5 Simulations

We have observed that for some hashtags, such as those relating to political subjects, users are particularly affected by multiple exposures before using them. We also know that the subgraphs on which political hashtags initially spread have high degrees and extensive clustering. To what extent do these aspects

intrinsically go together? Do these types of political hashtags spread effectively because of the close-knit network of the initial users? Are political subjects less likely to successfully spread on sparsely connected initial sets?

In this section, we try to obtain some initial insight into these questions through a simulation model — not only in the context of political hashtags but also in the context of the other categories. In particular, we develop a model that naturally complements the process used to calculate the  $p(k)$  functions. We perform simulations of this model using the measured  $p(k)$  functions and a varying number of the first users who used each hashtag on the actual influence network. Additionally, we record the progression of the cascade and track its spread through the network. By trying the  $p(k)$  curve of a hashtag on the initial sets of other hashtags, and by varying the size of the initial sets, we can gain insight into the factors that lead to wide-spreading cascades.

### 6.5.1 The Simulated Model

We wish to simulate cascades using the measured  $p(k)$  curves, the underlying network of users, and in particular the observed subgraphs  $G_m$  of initial adopters. In this discussion, and in motivating the model, we refer to the moment at which a node adopts a hashtag as its *activation*. We operationalize the model implicit in the definition of the function  $p(k)$ , leading to the following natural simulation process on a graph  $G = (V, E)$ .

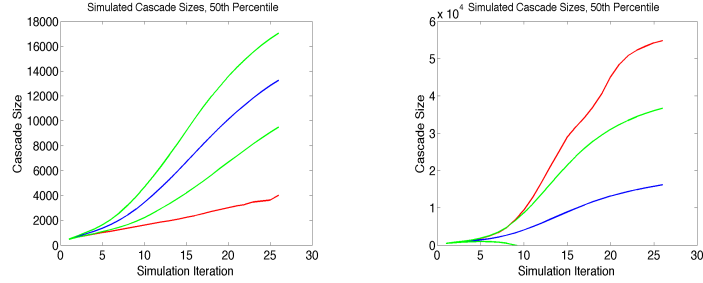
First, we activate all nodes in the starting set  $I$ , and mark them all as newly active. In a general iteration  $t$  (starting with  $t = 0$ ), we will have a currently active set  $A_t$  and a subset  $N_t \subseteq A_t$  of *newly active* nodes. (In the opening iteration,

we have  $A_0 = N_0 = I$ .) Newly active nodes have an opportunity to activate nodes  $u \in V - A_t$ , with the probabilities of success on  $u$  determined by the  $p(k)$  curve and the number of nodes in  $A_t - N_t$  who have already tried and failed to activate  $u$ .

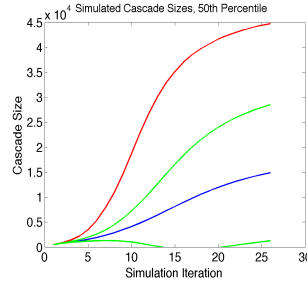
Thus, we consider each node  $u \in V - A_t$  that is a neighbor of at least one node in  $N_t$ , and hence will experience at least one activation attempt. Let  $k_t(u)$  be the number of nodes in  $A_t - N_t$  adjacent to  $u$ ; these are the nodes that have already tried and failed to activate  $u$ . Let  $\Delta_t(u)$  be the number of nodes in  $N_t$  adjacent to  $u$ . Each of these neighbors in  $N_t$  will attempt to activate  $u$  in sequence, and they will succeed with probabilities  $p(k_t(u) + 1), p(k_t(u) + 2), \dots, p(k_t(u) + \Delta_t(u))$ , since these are the success probabilities given the number of nodes that have already tried and failed to activate  $u$ . At the end, we define  $N_{t+1}$  to be the set of nodes  $u$  that are newly activated by the attempts in this iteration, and  $A_{t+1} = A_t \cup N_{t+1}$ .

## 6.5.2 Simulation Results

We simulate how a cascade that spreads according to the  $p(k)$  curve for some hashtag evolves when seeded with an initially active user sets of other hashtags. In total, there are 250,000 ( $p(k)$ , start set) hashtag combinations we examine. We additionally vary the size of the initially active set to be 100, 500, or 1,000 users. Since we want to study how a hashtag blossoms from being used by a few starting nodes to a large number of users, we must be careful about how we select the size of our starting sets. We believe that these initial set sizes capture the varying topology observed in Section 6.4 and are not too large as to guarantee wide-spreading cascade. For 100 and 500 starting nodes we run five



(a) Celebrity vs. random  $p(k)$  curves, celebrity start sets (b) Political vs. random start sets, political  $p(k)$  curves.



(c) Idiom vs. random start sets, idiom  $p(k)$  curves.

Figure 6.6: Validating Category Differences: The median cascade sizes for three different categories. In (a) we randomize over the  $p(k)$  curves and show that celebrity  $p(k)$  curves don't perform as well as random  $p(k)$  curves on celebrity start sets. Figures (b) and (c) illustrate the strength of the starting sets for political and idiom hashtags compared to random start sets. All starting sets consist of 500 users.

simulations on each  $(p(k), \text{start set})$  pair, and for 1,000 starting nodes we run only two simulations.

The simulation is instrumented at each iteration; we record the size of the cascade, the number of nodes influenced by active users, and the number of inactive users influenced by active users. Furthermore, each simulation runs for at most 25 iterations. We found that this number of iterations was large enough to observe interesting variation in cascade sizes yet still be efficiently simulated.

We calculate the mean and the 5th, 10th, ..., 95th percentiles of cascade sizes after each iteration. For each category, we measure these twenty measures based on all of the simulations where the  $p(k)$  hashtag and the starting set hashtag are both chosen from the category. We then compare these measurements to the results when a random set of hashtags is used to decide the  $p(k)$  curve, the starting set, or both the  $p(k)$  curve and the starting set. The cardinality of this random set is the same as the number of hashtags in the category. We sample these random choices 10,000 times to estimate the distribution of these measured features.

Using these samples, we test the measurements for statistical significance. In particular, we look at how the ‘category’ cascades (those in which both hashtag choices are from the category set) compare to cascades in which the  $p(k)$  curve or starting set hashtags were chosen randomly. In all of the following figures, the red line indicates the value of the measurements over the set of simulations in which  $p(k)$  curve and the start set come from category hashtags. The blue line is the average feature measurement over the random choices, and the green lines specify two standard deviations from the mean value. The cascade behavior of a category is statistically significant with respect to one of the measured features when most of the red curve lies outside of the region between the two green curves.

We compare how the  $p(k)$  curves for a category perform on start sets from the same category and on random start sets. We additionally evaluate how random  $p(k)$  curves and category  $p(k)$  curves perform on category start sets. In general, categories either performed below or above the random sets in both of these measures. Some particular observations are

- Celebrities and Games: Compared to random starting sets, we find that

start sets from these categories generate smaller cascades when the  $p(k)$  curves are chosen from their respective categories. This difference is statistically significant.

- Political and Idioms: These categories'  $p(k)$  curves and start sets perform better than a random choice. This is especially true for the smaller cascades (5 - 30th percentiles).
- Music: This category is interesting because the music  $p(k)$  curves perform better than random  $p(k)$  curves on music starting sets, music  $p(k)$  curves perform better on random starting sets than on music starting sets, regardless of the number of initially active users. This is the only category in which the  $p(k)$  and start set 'goodness' differs.
- Movies, Sports, and Technology: These categories don't exhibit particularly strong over or underperformance compared a random choice of  $p(k)$  hashtags and starting set hashtags.

## 6.6 Discussion

By studying the ways in which an individual's use of widely-adopted Twitter hashtags depends on the usage patterns of their network neighbors, we have found that hashtags of different types and topics exhibit different mechanics of spread. These differences can be analyzed in terms of the probabilities that users adopt a hashtag after repeated exposure to it, with variations occurring not just in the absolute magnitudes of these probabilities but also in their rate of decay. Some of the most significant differences in hashtag adoption provide intriguing confirmation of sociological theories developed in the off-line world. In partic-

ular, the adoption of politically controversial hashtags is especially affected by multiple repeated exposures, while such repeated exposures have a much less important marginal effect on the adoption of conversational idioms.

This extension of information diffusion analysis, taking into account sources of variation across topics, opens up a variety of further directions for investigation. First, the process of diffusion is well-known to be governed both by influence and also by homophily — people who are linked tend to share attributes that promote similarities in behavior. Recent work has investigated this interplay of influence and homophily in the spreading of on-line behaviors [6, 27, 7, 72]; It would be interesting to look at how this varies across topics and categories of information as well — it is plausible, for example, that the joint mention of a political hashtag provides stronger evidence of user-to-user similarity than the analogous joint mention of hashtags on other topics, or that certain conversational idioms (those that are indicative of shared background) are significantly better indicators of similarity than others. There has also been work on the temporal patterns of information diffusion — the rate over time at which different pieces of information are adopted [28, 62, 82, 90, 126]. In this context there have been comparisons between the temporal patterns of expected versus unexpected information [28] and between different media such as news sources and blogs [82]. Our analysis here suggests that a rich spectrum of differences may exist across topics as well.

Finally, we should emphasize one of our original points, that the phenomena we are observing are clearly taking place in aggregate: it is striking that, despite the many different styles in which people use a medium like Twitter, sociological principles such as the complex contagion of controversial topics can still



be observed at the population level. Ultimately, it will be interesting to pursue more fine-grained analyses as well, understanding how patterns of variation at the level of individuals contribute to the overall effects that we observe.

## CHAPTER 7

### INFLUENCE AND PASSIVITY IN SOCIAL MEDIA

The explosive growth of Social Media has provided millions of people the opportunity to create and share content on a scale barely imaginable a few years ago. Massive participation in these social networks is reflected in the countless number of opinions, news and product reviews that are constantly posted and discussed in social sites such as Facebook, Digg and Twitter, to name a few. Given this widespread generation and consumption of content, it is natural to target one's messages to highly connected people who will propagate them further in the social network. This is particularly the case in Twitter, which is one of the fastest growing social networks on the Internet, and thus the focus of advertising companies and celebrities eager to exploit this vast new medium. As a result, ideas, opinions, and products compete with all other content for the scarce attention of the user community. In spite of the seemingly chaotic fashion with which all these interactions take place, certain topics manage to get an inordinate amount of attention, thus bubbling to the top in terms of popularity and contributing to new trends and to the public agenda of the community. How this happens in a world where crowdsourcing dominates is still an unresolved problem, but there is considerable consensus on the fact that two aspects of information transmission seem to be important in determining which content receives attention.

One aspect is the popularity and status of given members of these social networks, which is measured by the level of attention they receive in the form of followers who create links to their accounts to automatically receive the content they generate. The other aspect is the influence that these individuals wield,

which is determined by the actual propagation of their content through the network. This influence is determined by many factors, such as the novelty and resonance of their messages with those of their followers and the quality and frequency of the content they generate. Equally important is the passivity of members of the network which provides a barrier to propagation that is often hard to overcome. Thus gaining knowledge of the identity of influential and least passive people in a network can be extremely useful from the perspectives of viral marketing, propagating one's point of view, as well as setting which topics dominate the public agenda.

In this chapter, we analyze the propagation of web links on Twitter over time to understand how attention to given users and their influence is determined. We devise a general model for influence using the concept of passivity in a social network and develop an efficient algorithm similar to the HITS algorithm [68] to quantify the influence of all the users in the network. Our influence measure utilizes both the structural properties of the network as well as the diffusion behavior among users. The influence of a user thus depends not only on the size of the influenced audience, but also on their passivity. This differentiates our measure of influence from earlier ones, which were primarily based on individual statistical properties such as the number of followers or retweets [23].

We have shown through extensive evaluation that this influence model outperforms other measures of influence such as PageRank, H-index, the number of followers and the number of retweets. In addition, it has good predictive properties in that it can forecast in advance the upper bound on the number of clicks a URL can get. We have also presented case studies showing the top influential users uncovered by our algorithm. An important conclusion from the

results is that the correlation between popularity and influence is quite weak, with the most influential users not necessarily being the ones with the highest popularity. Additionally, when we considered nodes with high passivity, we found the majority of them to be spammers and robot users. This demonstrates the applicability of our algorithm to automatic user categorization and filtering of online content.

## 7.1 Related work

The study of information and influence propagation in social networks has been particularly active for a number of years in fields as disparate as sociology, communication, marketing, political science and physics. Earlier work focused on the effects that scale-free networks and the affinity of their members for certain topics had on the propagation of information [131]. Others discussed the presence of key influentials [30, 43, 4, 129] in a social network, defined as those who are responsible for the overall information dissemination in the network. This research highlighted the value of highly connected individuals as key elements in the propagation of information through the network.

Huberman et al. [59] studied the social interactions on Twitter to reveal that the driving process for usage is a sparse hidden network underlying the friends and followers, while most of the links represent meaningless interactions. Jansen et al. [61] have examined Twitter as a mechanism for word-of-mouth advertising. They considered particular brands and products and examined the structure of the postings and the change in sentiments. Galuba et al. [34] propose a propagation model that predicts, which users will tweet about which URLs based on the history of past user activity.

There have also been earlier studies that focused on social influence and propagation. Agarwal et al. [4] have examined the problem of identifying influential bloggers in the blogosphere. They discovered that the most influential bloggers were not necessarily the most active. Aral et al [7] have distinguished the effects of homophily from influence as motivators for propagation. As to the study of influence within Twitter, Cha et al. [23] have performed a comparison of three different measures of influence - indegree, retweets and user mentions. They discovered that while retweets and mentions correlated well with each other, the indegree of users did not correlate well with the other two measures. Based on this, they hypothesized that the number of followers may not be a good measure of influence. On the other hand, Weng et al [129] have proposed a topic-sensitive PageRank measure for influence in Twitter. Their measure is based on the fact that they observed high reciprocity among follower relationships in their dataset, which they attributed to homophily. However, other work [23] has shown that the reciprocity is low overall in Twitter and contradicted the assumptions of this work.

## 7.2 Graph Construction

Twitter provides a Search API for extracting tweets containing particular keywords. To obtain the dataset for this study, we continuously queried the Twitter Search API for a period of 300 hours starting on 10 Sep 2009 for all tweets containing the string `http`. This allowed us to acquire a complete stream of all the tweets that contain URLs. We estimated the 22 million we accumulated to be 1/15th of the entire Twitter activity at that time. From each of the accumulated tweets, we extracted the URL mentions. Each of the unique 15 million URLs in

the dataset was then checked for valid formatting and the URLs shortened via the services such as `bit.ly` or `tinyurl.com` were expanded into their original form by following the HTTP redirects. For each encountered unique user ID, we queried the Twitter API for metadata about that user and in particular the user's followers and followees. The end result was a dataset of timestamped URL mentions together with the complete social graph for the users concerned.

**User graph.** The user graph contains those users whose tweets appeared in the stream, i.e., users that during the 300 hour observation period posted at least one public tweet containing a URL. The graph does not contain any users who do not mention any URLs in their tweets or users that have chosen to make their Twitter stream private.

For each newly encountered user ID, the list of followed users was only fetched once. Our dataset does not capture the changes occurring in the user graph over the observation period.

### 7.3 The IP Algorithm

**Evidence for passivity.** The users that receive information from other users may never see it or choose to ignore it. We have quantified the degree to which this occurs on Twitter (Fig. 7.1). An average Twitter user retweets only one in 318 URLs, which is a relatively low value. The retweeting rates vary widely across the users and the small number of the most active users play an important role in spreading the information in Twitter. This suggests that the level of user passivity should be taken into account for the information spread models to be accurate.

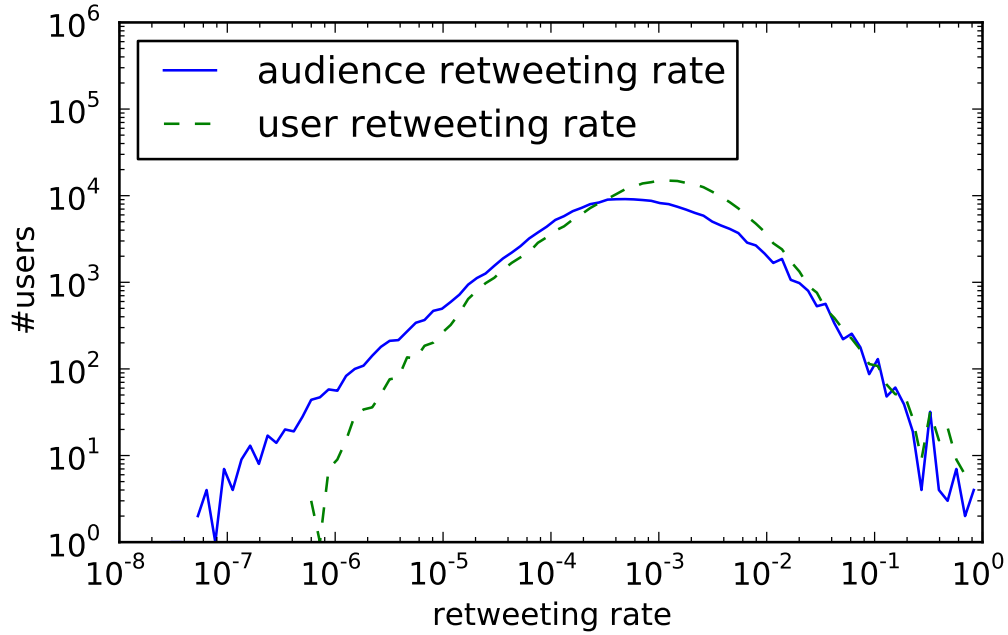


Figure 7.1: Evidence for the Twitter user passivity. We measure passivity by two metrics: 1. the user retweeting rate and 2. the audience retweeting rate. The *user retweeting rate* is the ratio between the number of URLs that user  $i$  decides to retweet to the total number of URLs user  $i$  received from the followed users. The *audience retweeting rate* is the ratio between the number of user  $i$ 's URLs that were retweeted by  $i$ 's followers to the number of times a follower of  $i$  received a URL from  $i$ .

**Assumptions.** Twitter is used by many people as a tool for spreading their ideas, knowledge, or opinions to others. An interesting and important question is whether it is possible to identify those users who are very good at spreading their content, not only to those who choose to follow them, but to a larger part of the network. It is often fairly easy to obtain information about the pairwise influence relationships between users. In Twitter, for example, one can measure how much influence user A has on user B by counting the number of times B retweeted A. However, it is not very clear how to use the pairwise influence

information to accurately obtain information about the relative influence each user has on the whole network. To answer this question, we design an algorithm (IP) that assigns a relative *influence* score and a *passivity* score to every user. The passivity of a user is a measure of how difficult it is for other users to influence him. Since we found evidence that users on Twitter are generally passive, the algorithm takes into account the passivity of all the people influenced by a user, when determining the user's influence. In other words, we assume that the influence of a user depends on both the quantity and the quality of the audience she influences. In general, our model makes the following assumptions:

1. A user's influence score depends on the number of people she influences as well as their passivity.
2. A user's influence score depends on how dedicated the people she influences are. Dedication is measured by the amount of attention a user pays to a given one as compared to everyone else.
3. A user's passivity score depends on the influence of those who she's exposed to but not influenced by.
4. A user's passivity score depends on how much she rejects other user's influence compared to everyone else.

**Operation.** The algorithm iteratively computes both the passivity and influence scores simultaneously in the following way:

Given a weighted directed graph  $G = (N, E, W)$  with nodes  $N$ , arcs  $E$ , and arc weights  $W$ , where the weights  $w_{ij}$  on arc  $e = (i, j)$  represent the ratio of influence that  $i$  exerts on  $j$  to the total influence that  $i$  attempted to exert on  $j$ , the IP algorithm outputs a function  $I : N \rightarrow [0, 1]$ , which represents the node's relative



influence on the network, and a function  $P : N \rightarrow [0, 1]$  which represents the node's relative passivity.

For every arc  $e = (i, j) \in E$ , we define the *acceptance rate* by  $u_{ij} = \frac{w_{i,j}}{\sum_{k:(k,j) \in E} w_{k,j}}$ .

This value represents the amount of influence that user  $j$  accepted from user  $i$  normalized by the total influence accepted by  $j$  from all users in the network.

The acceptance rate can be viewed as the dedication or loyalty user  $j$  has to user  $i$ . On the other hand, for every  $e = (j, i) \in E$  we define the *rejection rate* by

$v_{ji} = \frac{1 - w_{ji}}{\sum_{k:(j,k) \in E} (1 - w_{jk})}$ . Since the value  $1 - w_{ji}$  is the amount of influence that user

$i$  rejected from  $j$ , then the value  $v_{ji}$  represents the influence that user  $i$  rejected from user  $j$  normalized by the total influence rejected from  $j$  by all users in the network.

The algorithm is based on the following operations:

$$I_i \leftarrow \sum_{j:(i,j) \in E} u_{ij} P_j \quad (7.1)$$

$$P_i \leftarrow \sum_{j:(j,i) \in E} v_{ji} I_j \quad (7.2)$$

Each term on the right hand side of the above operations corresponds to one of the listed assumptions. In operation 1, the term  $P_j$  corresponds to assumption 1 and the term  $u_{ij}$  corresponds to assumption 2. In operation 2, the term  $I_j$  corresponds to assumption 3 and the term  $v_{ji}$  corresponds to assumption 4. The *Influence-Passivity algorithm* (Algorithm 1) takes the graph  $G$  as the input and computes the influence and passivity for each node in  $m$  iterations.

The IP algorithm is similar to the HITS algorithm for finding authoritative web pages and hubs that link to them [68]. The passivity score corresponds to the authority score, and the influence corresponds to hub score. However, IP

---

**Algorithm 1:** The Influence-Passivity (IP) algorithm

---

```
 $I_0 \leftarrow (1, 1, \dots, 1) \in \mathbf{R}^{|N|};$   
 $P_0 \leftarrow (1, 1, \dots, 1) \in \mathbf{R}^{|N|};$   
for  $i = 1$  to  $m$  do  
    Update  $P_i$  using operation (2) and the values  $I_{i-1}$ ;  
    Update  $I_i$  using operation (1) and the values  $P_i$ ;  
    for  $j = 1$  to  $|N|$  do  
         $I_j = \frac{I_j}{\sum_{k \in N} I_k};$   
         $P_j = \frac{P_j}{\sum_{k \in N} P_k};$   
    end  
end  
Return  $(I_m, P_m);$ 
```

---

is different from HITS in that it operates on a weighted graph and it takes into account other properties of the network such as those referred to as "acceptance rate" and "rejection rate."

**Generating the input graph.** There are many ways of defining the influence graph  $G = (N, E, W)$ . We construct it by taking into account retweets and the follower graph in the following way: The nodes are users who tweeted at least 3 URLs. The arc  $(i, j)$  exists if user  $j$  retweeted a URL posted by user  $i$  at least once. The arc  $e = (i, j)$  has weight  $w_e = \frac{S_{ij}}{Q_i}$  where  $Q_i$  is the number of URLs that  $i$  mentioned and  $S_{ij}$  is the number of URLs mentioned by  $i$  and retweeted by  $j$ .

## 7.4 Evaluation

### 7.4.1 Computations

Based on the obtained dataset we generate the weighted graph using the method described in section 7.3. The graph consists of approximately 450k nodes and 1 million arcs with mean weight of 0.07, and we use it to compute the PageRank, influence and passivity values for each node. The Influence-Passivity algorithm converges to the final values in tens of iterations (Fig. 7.2).

**PageRank.** The PageRank algorithm has been widely used to rank web pages as well as people based on their authority and influence [13]. In order to compare it with the results from the IP algorithm, we compute PageRank on the weighted graph  $G = (N, E, W)$  with a small change. First, since the arcs  $e = (i, j) \in E$  indicate that user  $i$  exerts some influence on user  $j$  then we invert all the arcs before running PageRank on the graph while leaving the weights intact. In other words, we generate a new graph  $G' = (N', E', W')$  where  $N' = N$ ,  $E' = \{(i, j) : (j, i) \in E\}$ , and for each  $(i, j) \in E'$  we define  $w'_{ij} = w_{ji}$ . This generates a new graph  $G'$  analogous to  $G$  but where the influenced users point to their influencers. Second, since the graph  $G'$  is weighted we assume that when the random surfer of the PageRank algorithm is currently at the node  $i$ , she chooses to visit node  $j$  next with probability  $\frac{w'_{ij}}{\sum_{k:(i,k) \in E'} w'_{ik}}$ .

**The Hirsch Index.** The Hirsch index (or H-index) is used in the scientific community in order to measure the productivity and impact of a scientist. A scientist has index  $h$  if he has published  $h$  articles which have been cited at least  $h$  times each. It has been shown that the H-index is a good indicator of whether

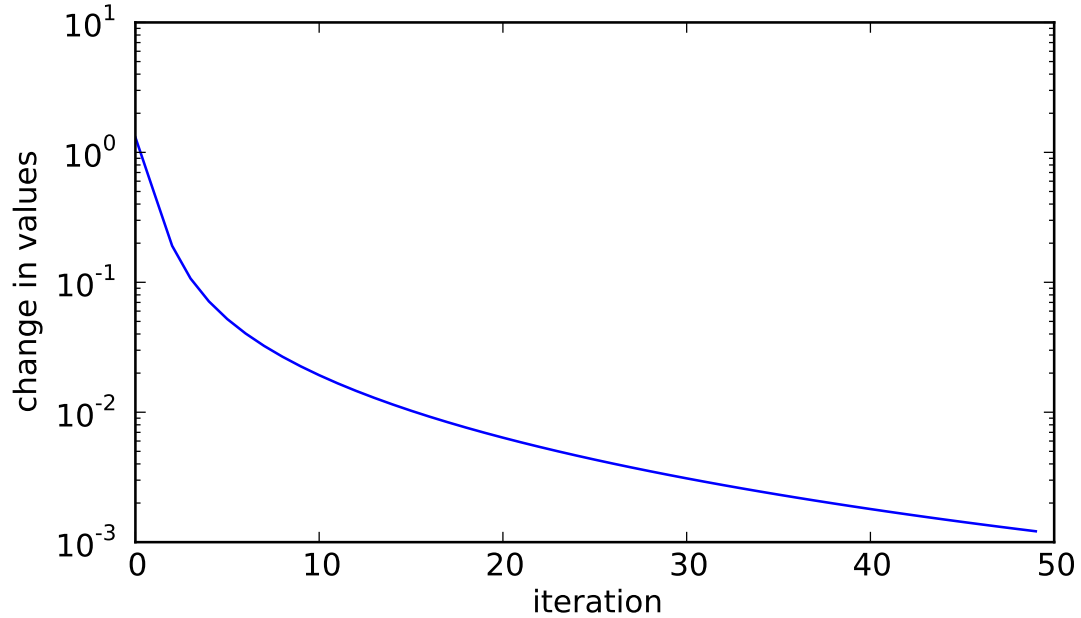
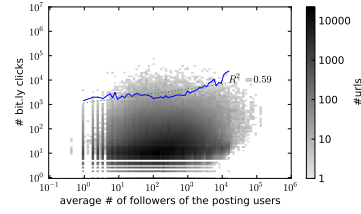


Figure 7.2: IP-algorithm convergence. In each iteration we measure the sum of all the absolute changes of the computed influence and passivity values since the previous iteration

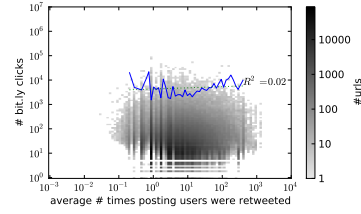
a scientist has had high achievements such as getting the Nobel prize [55]. Analogously, in Twitter, a user has index  $h$  if  $h$  of his URL posts have been retweeted at least  $h$  times each.

#### 7.4.2 Influence as a correlate of attention

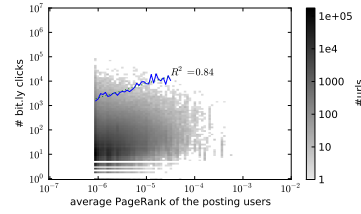
Any measure of influence is necessarily a subjective one. However, in this case, a good measure of influence should have a high predictive power on how well the URLs mentioned by the influential users attract attention and propagate in the social network. We would expect the URLs that highly influential users



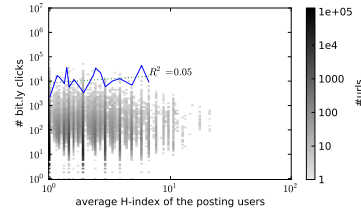
(a) Average number of followers vs. number of clicks on URLs



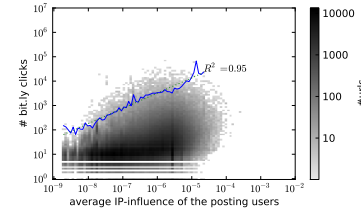
(b) Average number of times users were retweeted vs. number of clicks on URLs



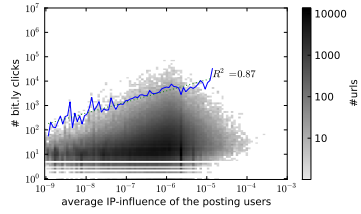
(c) Average user PageRank vs. number of clicks on URLs



(d) Average user H-index vs. number of clicks on URLs



(e) Average user IP-influence vs. number of clicks on URLs, using the retweet graph as input



(f) Average user IP-influence vs. number of clicks on URLs, using the co-mention graph as input

Figure 7.3: We consider several user attributes: the number of followers, the number of times a user has been retweeted, the user's PageRank, H-index and IP-influence. For each of the 3.2M Bit.ly URLs we compute the average value of a user's attribute among all the users that mentioned that URL. This value becomes the  $x$  coordinate of the URL-point; the  $y$  coordinate is the number of clicks on the Bit.ly URL. The density of the URL-points is then plotted for each of the four user attributes. The solid line in each figure represents the 99.9th percentile of Bit.ly clicks at a given attribute value. The dotted line is the linear regression fit for the solid line with the fit's  $R^2$  and slope displayed beside it.

propagate to attract a lot of attention and user clicks. Thus, a viable estimator of attention is the number of times a URL has been accessed.

**Click data.** Bit.ly is a URL shortening service that for each shortened URL keeps track of how many times it has been accessed. There are 3.2M unique Bit.ly URLs in the tweets from our dataset. We have queried the Bit.ly API for the number of clicks the service has registered on each URL.

A URL may be shortened by a user who has a Bit.ly account. Each such shortening is assigned a unique per-user Bit.ly URL. To account for that we took the “global clicks” number returned by the API instead of the “user clicks” numbers. The “global clicks” number sums the clicks across all the Bit.ly shortenings of a given URL and across all the users.

**URL traffic Prediction.** Using the URL click data, we take several different user attributes and test how well they can predict the attention the URLs posted by the users receive (Fig. 7.3). It is important to note that none of the influence measures are capable of predicting the exact number of clicks. The main reason for this is that the amount of attention a URL gets is not only a function of the influence of the users mentioning it, but also of many other factors including the virality of the URL itself and more importantly, whether the URL was mentioned anywhere outside of Twitter, which is likely to be the biggest source of unpredictability in the click data.

The wide range of factors potentially affecting the Bit.ly clicks may prevent us from predicting their number accurately. However, the upper bound on that number can to a large degree be predicted. To eliminate the outlier cases, we examined how the 99.9<sup>th</sup> percentile of the clicks varied as the measure of influence

increased.

**Number of followers.** The most readily available and often used by the Twitterers measure of influence is the number of followers a user has. As the Figure 7.3(a) shows, the number of followers of an average poster of a given URL is a relatively weak predictor of the maximum number of clicks that the URL can receive, with an  $R^2$  value of 0.59.

**Number of retweets.** When users post URLs, their posts might be retweeted by other users. Each retweet explicitly credits the original poster of the URL (or the user from whom the retweeting user heard about the URL). The number of times a user has been credited in a retweet has been assumed to be a good measure of influence [23]. However, Figure 7.3(b) shows that the number of times a user has been retweeted in the past is an extremely poor predictor of the maximum number of clicks the URLs posted by that user can get.

**The Hirsch Index.** Figure 7.3(d) shows that despite the fact that in the scientific community the H-index is used as a good predictor of scientific achievements, in Twitter, it has very low correlation with URL popularity ( $R^2$  of 0.05). This may reflect the fact that attention in the scientific community plays a symmetric role, since those who pay attention to the work of others also seek it from the same community. Thus, citations play a strategic role in the successful publishing of papers, since the expectation of authors is that referees and authors will demand attention to their work and those of their colleagues. Within Social Media such symmetry does not exist and thus the decision to forward a message to the network lacks this particularly strategic value.

**PageRank.** Figure 7.3(c) shows that the average PageRank of those who

tweet a certain URL is a much better predictor of the URL's traffic than the average number of followers, retweets, or Hirsch index. The reason for the improvement could be explained by the fact that PageRank takes into account structural properties of the graph as opposed to individual measures of the users. However, figure 7.3(c) also shows that IP influence is a better indicator of URL popularity than PageRank. One of the main differences between the IP algorithm and PageRank is that the IP algorithm takes into account the passivity of the people a user influences and PageRank does not. This suggests that influencing users who are difficult to influence, as opposed to simply influencing many users, has a positive impact on the eventual popularity of the message that a user tweets.

**IP-Influence score.** As we can see in Figure 7.3(e), the average IP-influence of those who tweeted a certain URL can determine the maximum number of clicks that a URL will get with good accuracy, achieving an  $R^2$  score of 0.95. Since the URL clicks are never considered by the IP algorithm to compute the user's influence, the fact that we find a very clear connection between average IP-influence and the eventual popularity of the URLs (measured by clicks) serves as an unbiased evaluation of the algorithm and demonstrates the utility of IP-influence. For example, as we can see in Figure 7.3(e), given a group of users having very large average IP-influence scores who post a URL we can estimate, with 99.9% certainty, that this URL will not receive more than 100,000 clicks. On the other hand, if a group of users with very low average IP-influence score post the same URL we can estimate, with 99.9% certainty that the URL will not receive more than 100 clicks.

Furthermore, figure 7.4 shows that a user's IP-influence is not well correlated with the number of followers she has. This reveals interesting implications



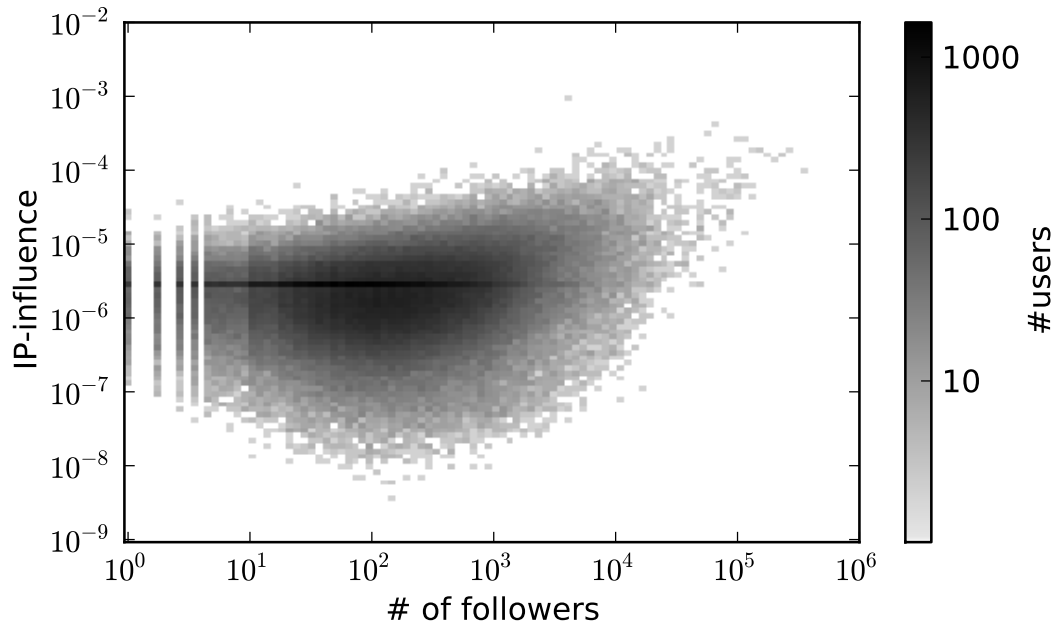


Figure 7.4: For each user we place a user-point with IP-influence as the  $y$  coordinate and the  $x$  coordinate set to the number of user's followers. The density of user-points is represented in grayscale. The correlation between IP-influence and #followers is 0.44.

about the relationship between a person's popularity and the influence she has on other people. In particular, it shows that having many followers on Twitter does not directly imply the power to influence them to click on a URL.

In the above experiments, we have used the average number of followers, retweets, PageRank, H-Index, and IP-influence of the users who posted a URL to predict the URL's traffic. We examined other choices such as using the maximum number instead of the average, and obtained similar results.

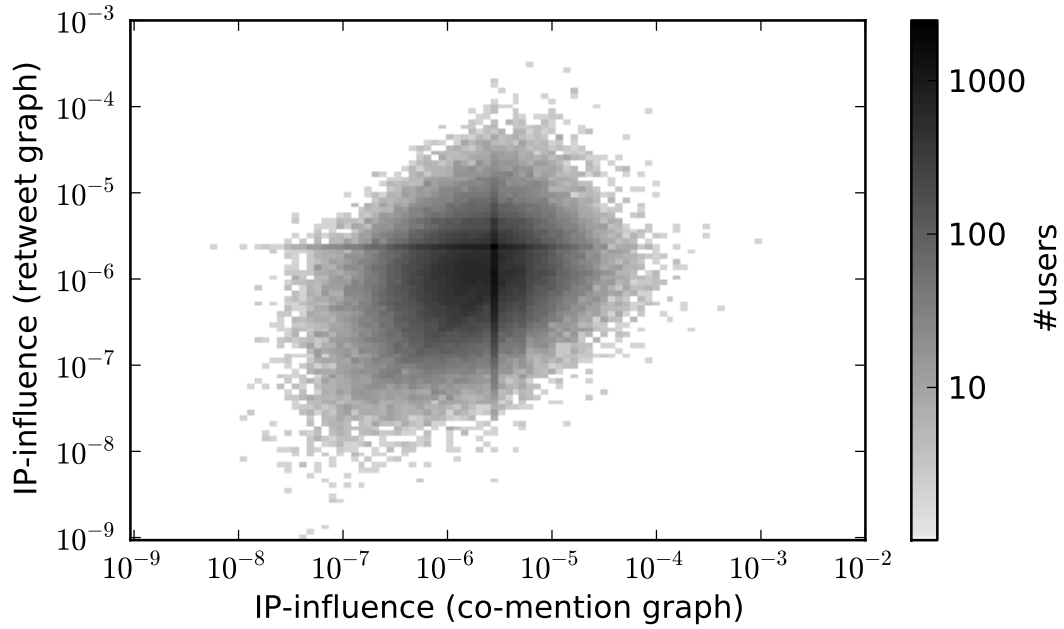


Figure 7.5: The correlation between the IP-influence values computed based on two inputs: the co-mention influence graph and the retweet influence graph. The correlation between the two influence values is 0.06.

## 7.5 IP Algorithm Adaptability

As mentioned earlier, there are many ways of defining a social graph in which the edges indicate pairwise influence. We have so far been using the graph based on which user retweeted which user (*retweet influence graph*). However, the explicit signals of influence such as retweets are not always available. One way of overcoming this obstacle is to use other, possibly weaker, signals of influence. In the case of Twitter, we can define an influence graph based on mentions of URLs without regard of actual retweeting in the following way.

**The co-mention graph.** The nodes of the *co-mention influence graph* are users

who tweeted at least three URLs. The edge  $(i, j)$  exists if user  $j$  follows user  $i$  and  $j$  mentioned at least one URL that  $i$  had previously mentioned. The edge  $e = (i, j)$  has weight  $w_e = \frac{S_{ij}}{F_{ij} + S_{ij}}$  where  $F_{ij}$  is the number of URLs that  $i$  mentioned and  $j$  never did and  $S_{ij}$  is the number of URLs mentioned by  $j$  and previously mentioned by  $i$ .

The resulting graph has the disadvantage that the edges are based on a much less explicit notion of influence than when based on retweets. Therefore the graph could have edges between users who do not influence each other. On the other hand, the retweeting conventions on Twitter are not uniform and therefore sometimes users who repost a URL do not necessarily credit the correct source of the URL with a retweet [12]. Hence, the influence graph based on retweets has potentially missing edges.

Since the IP algorithm has the flexibility of allowing any influence graph as input, we can compute the influence scores of the users based on the co-mention influence graph and compare with the results obtained from the retweet influence graph. As we can see in Figure 7.3(f), we find that the retweet graph yields influence scores that are better at predicting the maximum number of clicks a URL will obtain than the co-mention influence graph. Nevertheless, Figure 7.3(f) shows that the influence values obtained from the co-mention influence graph are still better at predicting URL traffic than other measures such as PageRank, number of followers, H-index or the total number of times a user has been retweeted. Furthermore, Figure 7.5 shows that the influence score based on both graphs do not correlate well, which suggests that considering explicit vs. implicit signals of influence can change the outcome of the IP algorithm, while at the same time maintaining its predictive value. In general, we find that the

mashable	Social Media Blogger
jokoanwar	Film Director
google	Google
aplusk	Actor Ashton Kutcher
syfy	Science Fiction Channel
smashingmag	Online Developer Magazine
micellemalkin	Conservative Commentator
theonion	News Satire Organization
rww	Tech/Social Media Blogger
breakingnews	News Aggregator

Table 7.1: Users with the most IP-influence (with at least 10 URLs posted in the period)

explicitness of the signal provided by the retweets yields slightly better results when it comes to predicting URL traffic, however, the influence scores based on co-mentions may surface a different set of potentially influential users.

## 7.6 Case Studies

As we mentioned earlier, one important application of the IP algorithm is ranking users by their relative influence. In this section, we present a series of rankings of Twitter users based on the influence, passivity, and number of followers.

**The most influential.** Table 1 shows the users with the most IP-influence in the network. We constrain the number of URLs posted to 10 to obtain this list, which is dominated by news services from politics, technology, and Social Media. These users post many links which are forwarded by other users, causing their influence to be high.

**The most passive.** Table 2 shows the users with the most IP-passivity in the

network. Passive users are those who follow many people, but retweet a very small percentage of the information they consume. Interestingly, robot accounts (which automatically aggregate keywords or specific content from any user on the network), suspended accounts (which are likely to be spammers), and users who post extremely often are among the users with the most IP-passivity. Since robots “attend” to all existing tweets and only retweet certain ones, the percentage of information they forward from other users is actually very small. This explains why the IP-algorithm assigns them such high passivity scores. This also highlights a new application of the IP-algorithm: automatic identification of robot users including aggregators and spammers.

redscarebot	Keyword Aggregator
drunk_bot	Suspended
tea_robot	Keyword Aggregator
condos	Listing Aggregator
wootboot	Suspended
raybeckerman	Attorney
hashphotography	Keyword Aggregator
charlieandsandy	Suspended
ms_defy	Suspended
rpattinsonbot	Keyword Aggregator

Table 7.2: Users with the most IP-passivity

**The least influential with many followers.** We have demonstrated that the amount of attention a person gets may not be a good indicator of the influence they have in spreading their message. In order to make this point more explicit, we show, in Table 3, some examples of users who are followed by many people but have relatively low influence. These users are very popular and have the attention of millions of people but are not able to spread their message very far. In most cases, their messages are consumed by their followers but not considered important enough to forward to others.

User name	Category	# followers Rank	IP-influence Rank
thatkevinsmith	Screen Writer	33	1000
nprpolitics	Political News	41	525
eonline	TV Channel	42	1008
marthastewart	Television Host	43	1169
nba	Sports	64	1041
davidgregory	Journalist	106	3630
nfl	Sports	110	2244
cbsnews	News Channel	114	2278
jdickerson	Journalist	147	4408
newsweek	News Magazine	148	756

Table 7.3: Users with many followers and low relative influence

**The most influential with few followers.** We are also able identify users with very low number of followers but high influence. Table 4 shows the users with the most influence who rank less than 100,000<sup>th</sup> in number of followers. We find that during the data collection period some of the users in this category ran very successful retweeting contests where users who retweeted their URLs would have the chance of winning a prize. Moreover, there is a group of users who post from `twitdraw.com`, a website where people can make drawings and post them on Twitter. Even though these users don't have many followers, their drawings are of very high quality and spread throughout Twitter reaching many people. Other interesting users such as local politicians and political cartoonists are also found in the list. The IP-influence measure surfaces interesting content posted by users who would otherwise be buried by popularity rankings such as number of followers.

User name	Category	# followers Rank	IP-influence Rank
cashcycle	Retweet Contest	153286	13
mobiliens	Retweet Contest	293455	70
jadermattos	Twitdraw	227934	134
_jaum_	Twitdraw	404385	143
robmillerusmc	Politician	147803	145
sitekulite	Twitdraw	423917	149
jesse_sublett	Musician	385265	151
cyberaurora	Tech News Website	446207	163
viveraxo	Twitdraw	458279	165
fireflower_	Political Cartoons	452832	195

Table 7.4: Users with very few followers but high relative influence

## 7.7 Discussion

**Influence as predictor of attention.** As we demonstrated in section 7.4, the IP-influence of the users is an accurate predictor of the upper bound on the total number of clicks they can get on the URLs they post. The input to the influence algorithm is a weighted graph, where the arc weights represent the influence of one user over another. This graph can be derived from the user activity in many ways, even in cases where explicit feedback in the form of retweets or “likes” is not available.

**Topic-based and group-based influence.** The Influence-Passivity algorithm can be run on a subgraph of the full graph or on the subset of the user activity data. For example, if only users tweeting about a certain topic are part of the graph, the IP-influence determines the most influential users in that topic. It is an open question whether the IP algorithm would be equally accurate at different graph scales.

**Content ranking.** The predictive power of IP-influence can be used for content

filtering and ranking in order to reveal content that is most likely to receive attention based on which users mentioned that content early on. Similarly, as in the case of users, this can be computed on a per-topic or per-user-group basis.

**Content filtering.** We have observed from our passivity experiments that highly passive users tend to be primarily robots or spammers. This leads to an interesting extension of this work to perform content filtering, limiting the tweets to influential users and thereby reducing spam in Twitter feeds.

**Influence dynamics.** We have computed the influence measures over a fixed 300-hour window. However, the Social Media are a rapidly changing, real-time communication platform. There are several implications of this. First, the IP algorithm would need to be modified to take into account the tweet timestamps. Second, the IP-influence itself changes over time, which brings a number of interesting questions about the dynamics of influence and attention. In particular, whether users with spikes of IP-influence are overall more influential than users who can sustain their IP-influence over time is an open question.

Given the mushrooming popularity of Social Media, vast efforts are devoted by individuals, governments and enterprises to getting attention to their ideas, policies, products, and commentary through social networks. But the very large scale of the networks underlying Social Media makes it hard for any of these topics to get enough attention in order to rise to the most trending ones. Given this constraint, there has been a natural shift on the part of the content generators towards targeting those individuals that are perceived as influential because of their large number of followers. In this chapter we show that the correlation between popularity and influence is weaker than it might be expected. This is a reflection of the fact that for information to propagate in a network,



individuals need to forward it to the other members, thus having to actively engage rather than passively read it and rarely act on it. Moreover, since our measure of influence is not specific to Twitter it is applicable to many other social networks. This opens the possibility of discovering influential individuals within a network which can on average have a further reach than others in the same medium, regardless of their popularity.

## **Part III**

**Interplay between Network**

**Evolution and Information**

**Diffusion**

## CHAPTER 8

### SOCIAL-TOPICAL AFFILIATIONS

As discussed in section 2, there have been many studies that investigate how edges in social networks form and how networks evolve [32, 101, 60, 73, 85]. Furthermore, many studies have looked at online information sharing, either on social tag systems [50, 57, 93, 106, 51] or on the mechanisms of information diffusion [114, 48, 66]. It is an interesting question whether the information held by individuals in the network could describe properties of the existing social network, and whether the social network itself could describe the properties of the information that diffuses on it. We could think of a system where each user in a network is tagged with a certain topic if information related to that topic passes through the user, or if there is evidence that the user is interested in the topic. For example, the user may have downloaded a movie about the topic. This would generate a topical affiliation system that could be useful in understanding the structure and evolution of the network itself. In this chapter, we aim to bridge both the social and informational aspects together and study the ways in which they are related to each other. Furthermore, we investigate the extent to which the relation between social and informational aspects can be useful to predict each other using basic features and a standard predictive model. To do so, we look at the intersection of two key structures of online social media – the set structure of topical affiliations and the graph structure of social networks. We aim to understand the interplay between the two, thereby understanding “the social structure of topics”, and the “topical structure of social.”

We begin by looking at the classical problem of link prediction in social net-

works. Many people have studied this problem and the approach has often been to look at the features of the existing network in order to predict future connections [88, 121, 117]. In this chapter, we instead use features exclusively related to the topical proximity of the users as well as the graph properties of these topics, and demonstrate how prediction models based on topical features can yield impressive prediction accuracy. This part of the chapter shows how the topical structure of the users in the network can inform our understanding of the structure of the network itself.

Next, we develop our main result, understanding the extent to which the structure of the network determines the topical information diffusion process. In particular, we are interested in whether the structure of the graph induced by the initial set of adopters of a certain topic can tell us something about the eventual popularity of the topic. Here, we are using “topic” in a loose sense, referring to a product, idea, or even a behavior. The idea that the speed and magnitude of adoption of products, ideas, and behaviors can be driven by “viral marketing” techniques has gained tremendous popularity over the past few years [15, 40, 109, 81, 111]. The premise is that one can utilize the edges of an existing social network as bridges for information to spread from person to person. A common question is what kind of topics will go viral in the future, and the focus of our study is to ask: how precisely is this related to graph structure? To shed light into this question, we test a predictive algorithm that takes as features properties of the induced subgraph of the early adopters of a topic, with the goal of predicting its eventual popularity.

We find that the structure of the early adopter graphs can indeed have predictive power about the popularity of the topic. Furthermore, we observe that

the relationship between the topological properties of the initial graphs and the topic's popularity is not always as expected. For example, popularity of a topic does not monotonically change with the number of social connections among the initial users. Instead, we find that the topic exhibits high future popularity when the number of connections is either very high or very low. This could come as a surprise since few connections among adopters of a topic could be perceived as the topic lacking "virality."

We use Twitter *hashtags* – labels that users include in their posts to indicate the topic of the message – to distinguish between different topics and the follower and @-message network as a proxy for social connections among the users. Because these networks are directed and the @-message network is weighted, we are able to compare the differences in our results when considering reciprocate and unreciprocated relationships, as well as strong and weak ties. Through our analysis we also discover that while strong and mutual ties are easier to predict, they are less useful than weaker directed ties when predicting hashtag popularity.

## 8.1 Dataset

The dataset used in this chapter consists of two main parts: hashtags and networks.

**Hashtags.** As we discussed in chapter 6, hashtags are a convention widely used by Twitter users is a tagging system where a user includes a single unlimited string proceeded by a "#" character. This string is referred as *hashtag* and it is meant to label the tweet so other people know what it is about, the

designate that it belongs to a particular conversation topic. For example: “*What a game last night between the Thunder and Grizzlies. Was up till 1am watching that triple over-time thriller. #NBA.*” We extract all the hashtags that have appeared in our dataset and users who have utilized at least one hashtag. Our data set contains a total of 7,305,414 hashtags and 5,513,587 users who utilized at least one hashtag. On average, each hashtag is used by 9.48 distinct users, while a user posted about 12.57 different hashtags.

**Graphs.** We obtained the follower/followee network from [77], which contained the list of people each person was following at crawl time. If user  $A$  follows user  $B$  we create the edge  $(A, B)$ . There are around 366 million edges among the users who have utilized at least one hashtag. We also use a second graph based on *@-messages*. This is the same graph used in chapters 6 and 5.

## 8.2 Topical features predict links

In this section, we ask to what extent the hashtags that an individual has used reveals their ties to other users in Twitter’s directed social graph, or their ties to other users via *@-communication*. By allowing hashtags to define user sets, we can view Twitter users as embedded in the set system of these hashtags. We begin by characterizing the features of the hashtag usage of two individuals that we wish to process. We then consider a prediction problem, where we are trying to predict the presence of an edge between arbitrary pairs of individuals. From this, we observe that the size of the smallest common hashtag that two users overlap on is a surprisingly informative predictor. Having observed this, we ask to what extent the graph structure of these smallest common hashtags

can be used to improve prediction accuracy, ultimately obtaining remarkably a capable predictive model.

### 8.2.1 Measuring topic distance

In order to approach this question, we must first summarize the hashtag usage similarity of two individuals into features that could plausibly serve as similarity/distance measures. Perhaps the most obvious measure of similarity is the number of hashtags that two users have in common. This measure is immediately problematic, since it does not distinguish between hashtags that are broadly adopted and those that have only been used by a handful of users. More appropriately, we consider features that relate to the frequency of the common hashtags in the broader population. For this, we consider the size of the smallest common hashtag, the size of the largest, the average size, and also two measures that aggregate the common overlap of the full sets: the sum distance and the Adamic-Adar distance [1].

Consider the following notation:

- Let  $u_1, \dots, u_N$  be the  $N$  users.
- Let  $h_1, \dots, h_M$  be the  $M$  hashtags.
- Let  $H(u_i)$ , be the set of hashtags used by users  $u_i$ .
- Let  $U(h_j)$  be the users who used hashtag  $h_j$ .

This is the structural information we aim to process. The features of the common hashtags between users that we consider in the work are:

- The number of hashtags in common,  
 $|H(u_i) \cap H(u_j)|.$
- The size of the smallest common hashtag,  
 $\min_{h \in H(u) \cap H(v)} |U(h)|.$
- The size of the largest common hashtag,  
 $\max_{h \in H(u) \cap H(v)} |U(h)|.$
- The average size of the common hashtags,  
 $\frac{1}{|H(u) \cap H(v)|} \sum_{h \in H(u) \cap H(v)} |U(h)|.$
- The sum of the inverse sizes,  
 $\sum_{h \in H(u) \cap H(v)} 1/|U(h)|.$
- The Adamic-Adar distance,  
 $\sum_{h \in H(u) \cap H(v)} 1/\log |U(h)|.$

The size of the smallest common hashtag is an intuitively attractive measure: it captures the extent to which the conversations two persons share are unique or not. In [67], Kleinberg studied social networks where individuals were viewed as embedded in a set system (much like our hashtag set system) and an individual  $u$  was linked to an individual  $v$  with probability proportional to  $d(u, v)^{-a}$ , where  $d(u, v)$  is the size of the smallest common set. Kleinberg showed that decentralized greedy routing with regard to this measure takes polylogarithmic time if and only if  $a = 1$ . Beyond studying the performance of the minimum common hashtag size in a predictive setting, we therefore also investigate the extent to which such an inverse proportional dependence holds true in our setting.

The size of the largest common hashtag is an intuitively poor feature for predicting links, and we include it here specifically as a control of sorts, showing



that not all features of the common hashtags are informative. Very commonly, the largest common hashtag that users overlap on is one of a few extremely popular hashtags, for example #musicmonday, #ff, or #fail. In choosing to study the sum distance and Adamic-Adar distance — a measure introduced for studying the similarity of web-based social networks derived from common homepage content [1] — we allow ourselves to consider similarity measures using all common sets.

### 8.2.2 Predictive model

Given the features defined above, we now investigate how well the topical overlap represented in the common hashtags allow us to predict the presence of links. We formulate this task as a balanced classification task: given a set of 100,000 users that coincide on some hashtag, where 50,000 are disconnected pairs and 50,000 are connected, what sort of prediction accuracy can we obtain, compared to a naive 50% baseline?

We consider only user pairs that coincide on some hashtag because performing classification against completely arbitrary user pairs (where only 4.9% of hashtag-usinguser pairs coincide on some hashtag) would have been overly generous to our classifier, since arbitrary users rarely coincide on any hashtags, and the coincidence rates between connected users is understandably going to be higher. In fact, 35% of user pairs where one user follows the other coincide on some hashtag, and fully 78% coincide when comparing user pairs where one person has @-messed the other person at least once.

We approach the problem using logistic regression with 10-fold cross-

Directed edges	Model Features	Follow	@ ≥ 1	@ ≥ 3	@ ≥ 5	@ ≥ 7	@ ≥ 9	@ ≥ 20
	All hashtag features	0.737	0.826	0.850	0.860	0.862	0.870	0.871
	# common HTs	0.713	0.781	0.798	0.800	0.804	0.809	0.816
	Smallest HT size	0.703	0.799	0.828	0.841	0.842	0.854	0.855
	Largest HT size	0.582	0.587	0.584	0.585	0.581	0.583	0.575
	Average HT size	0.589	0.662	0.683	0.702	0.697	0.723	0.720
	Sum distance	0.712	0.804	0.832	0.845	0.848	0.858	0.860
	Adamic-Adar distance	0.727	0.809	0.831	0.842	0.846	0.852	0.856
	Hashtag features + Edges	<b>0.766</b>	<b>0.863</b>	<b>0.889</b>	<b>0.921</b>	<b>0.940</b>	<b>0.949</b>	<b>0.976</b>
	Edges of smallest	0.647	0.790	0.816	0.827	0.865	0.872	0.886
Mutual edges	Model Features	Follow	@ ≥ 1	@ ≥ 3	@ ≥ 5	@ ≥ 7	@ ≥ 9	@ ≥ 20
	All hashtag features	0.762	0.827	0.868	0.869	0.868	0.867	0.866
	# common HTs	0.739	0.782	0.809	0.813	0.812	0.812	0.808
	Smallest HT size	0.715	0.803	0.849	0.853	0.852	0.852	0.856
	Largest HT size	0.576	0.562	0.590	0.583	0.574	0.569	0.548
	Average HT size	0.597	0.671	0.712	0.706	0.707	0.706	0.743
	Sum distance	0.725	0.808	0.854	0.857	0.856	0.856	0.860
	Adamic-Adar distance	0.751	0.807	0.850	0.854	0.852	0.852	0.849
	Hashtag features + Edges	<b>0.796</b>	<b>0.864</b>	<b>0.922</b>	<b>0.936</b>	<b>0.934</b>	<b>0.949</b>	<b>0.967</b>
	Edges of smallest	0.651	0.788	0.829	0.832	0.833	0.837	0.861

Table 8.1: Prediction accuracies for directed and mutual edges, as trained on the full set of hashtag features, individual hashtag features, and edge features. Accuracy was evaluated using 10-fold cross-validation on a balanced classification dataset.

validation. For all six features, in addition to a linear term, we also include the logarithm and the inverse of each value, allowing us to more robustly extract non-linear dependencies.

We consider the tasks of predicting follow edges, mutual follow edges, @-message edges, and mutual @-message edges. It is natural to view the number of @-edges as the strength of a tie, and we therefore also consider our classification task applied to @-edges thresholded on high @-message counts. The accuracies we obtain are shown in Table 8.2.2, where we report the performance of both a full model, considering all features under all transformations, and also models trained on a single feature set (where we include its two transformations).

We see that classification based on common hashtag usage exhibits powerful

prediction accuracy given its simplicity: for follow edges our accuracy is 74%, for @-edges it is 83%, and for strong @-edges (more than 20 messages), our accuracy is 87%. We achieve comparable performance when predicting mutual edge relationships. Beyond this general performance, we note that the size of the smallest common hashtag is a consistently accurate feature when considered alone, especially when trying to predict strong ties. As expected, the size of the largest common hashtag is least performative.

### 8.2.3 Predictive model with edges

The features we consider above do not extract anything about the graph structure of the induced subgraphs that each hashtag defines. Given the accuracy of models based solely on the size of the smallest common hashtag, here we chose to investigate to what extent the social structure of these smallest common hashtags — the most unique topic that two users have in common — can further improve our predictions.

One motivation for considering the graph structure is the observation, shown in Figure 8.1, that the edge density for similarly sized hashtags can differ considerably. Simply put, some hashtags — some topics — are much more ‘social’ than others. For social-topical contexts involving geography, this would translate to knowing that two people both visit a bar versus knowing that they both visit the same bank. Both may be equally popular locations to visit, but the bar is a decidedly more social environment, and visiting the same bar is more likely to predict social interaction than visiting the same bank.

When performing this classification, it is important to avoid inadvertently

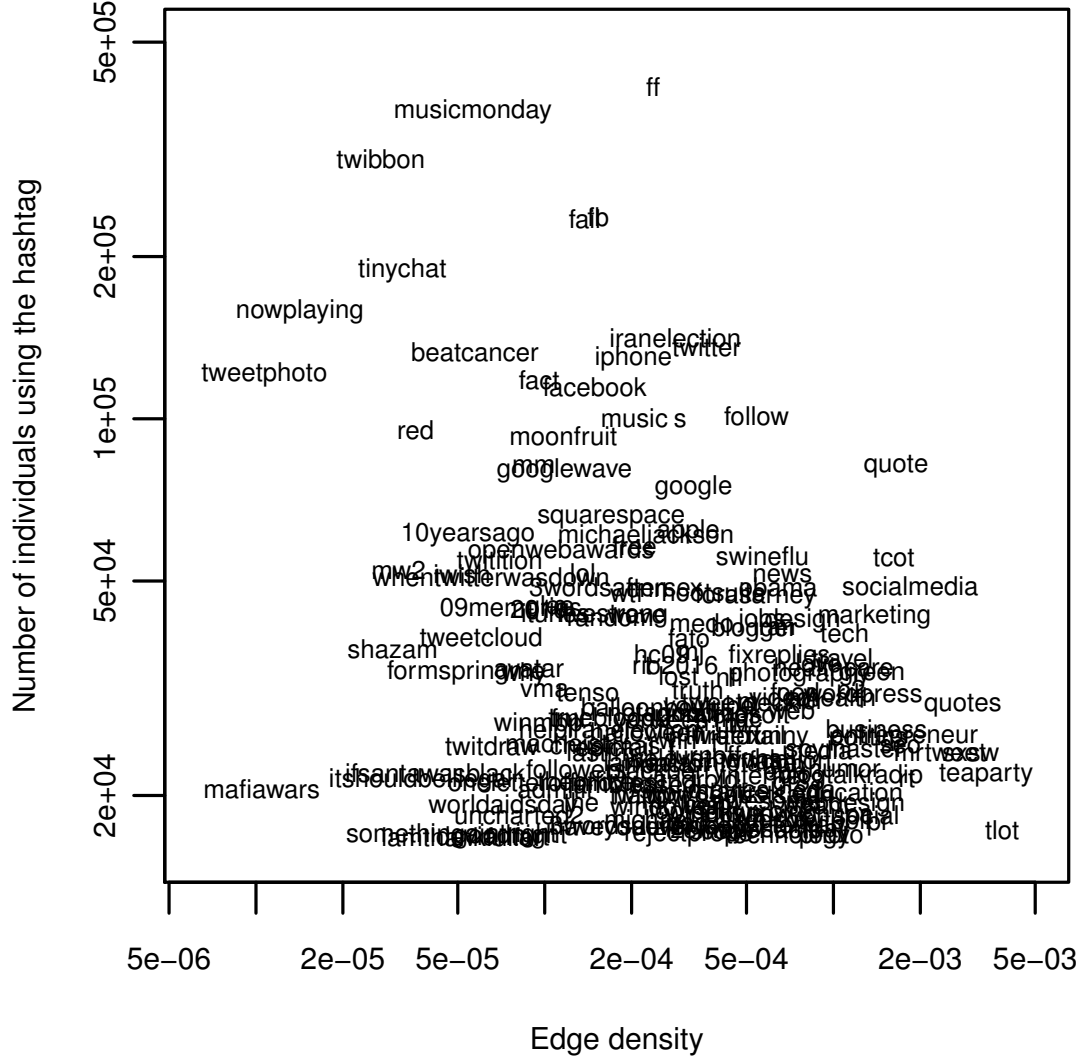


Figure 8.1: Edge density heterogeneity for the 100 most common hashtags in the dataset.

incorporating a circular reference whereby the link is directly present in the feature. Consider the problem of predicting an edge between a pair of users where the induced subgraph for their smallest common hashtag has two nodes and one edge. If we don't make this correction, we would know for certain that these two users are connected. We therefore let our edge count feature encode the number of edges present in the smallest common hashtag between users

other than the two users being considered.

By including the count of such edges appearing in the smallest common hashtag as a feature, where the type of the edges is the same as the type we are trying to predict, we see that our classification performance becomes remarkably accurate, demonstrating an accuracy of 76.6% when classifying follower relationships and 97.6% when classifying strongly tied @-edges.

## 8.2.4 Routing

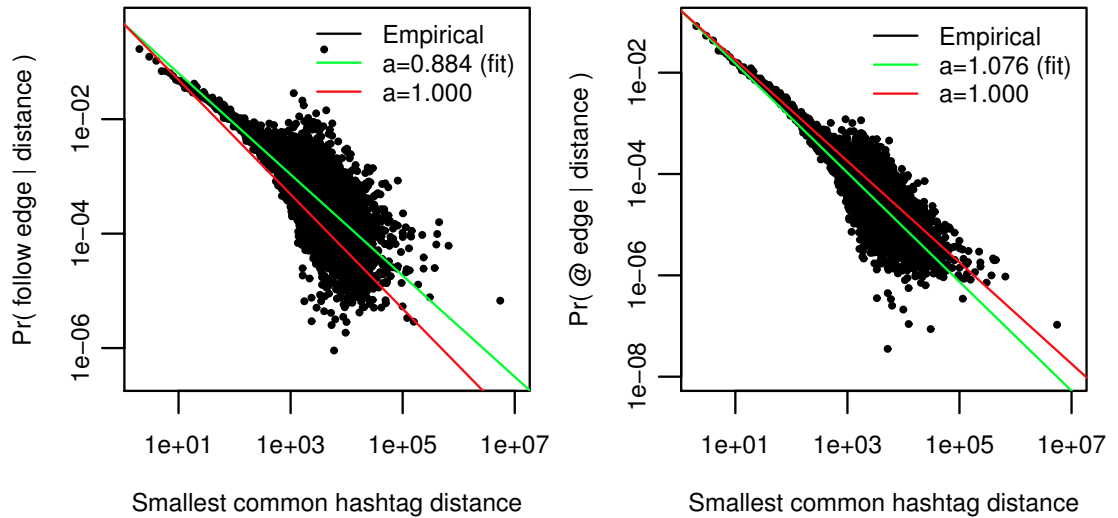


Figure 8.2: Linkage probability as a function of smallest common hashtag. (a) The probability of a given user following another user as a function of the size, and (b) the probability of a given user @-messaging another user as a function of size. Both figures are log-log scale.

In this section we briefly discuss whether the user-generated hashtag set system is amenable to greedy routing according to the smallest common set distance measure. From Figure 8.2, we see that aside from considerable hetero-

geneity, the probability of a link as a function of minimum common hashtag distance very much appears to obey an inverse power law. What In the case of @-communication (here thresholded on 3 messages), the probability of linkage appears to be moderately close to the  $a = 1$  necessary for efficient decentralized routing.

What would greedy routing via hashtags on Twitter mean? In practice, this would mean that if user  $u$  was trying to route a message to user  $v$  via the Twitter social graph, using only knowledge of what hashtags user  $v$  had used, they would greedily pass the message to their graph neighbor who was closest to  $v$  in the above defined ‘minimal common hashtag’ distance, and instruct that neighbor to pass the message along using the same greedy heuristic. If the social graph were perfectly embedded in the set system with the required structure, specifically with  $a = 1$ , then the number of steps needed to route the message would be only polylogarithmic in the number of Twitter users, which should be considered surprisingly efficient.

For the purposes of routing information on Twitter, this procedure would not be of practical interest, but observing the presence of this structure does however have serious implications for understanding social networks in a much broader sense. To the extent that follow and @-communication behavior on Twitter reflects social structure in society at large, observing this structure becomes a statement about how to find people in society based only on interests: if you are looking to meet with a particular person, and all you know are their ‘interests’ in some general sense, this result dictates that you should be able to efficiently approach them through your social network by navigating your search greedily with regard to these interests. Previous studies of online social

networks have suggested that greedy routing with regard to geographic distance is very nearly successful in this sense [103], but to our knowledge this is the first study to empirically investigate greedy routing on social networks purely based on interests.

Computing these link probabilities required evaluating the distances of both the linked users as well as the distances of all non-linked pairs of users, a significant hurdle. Because there are over 5 million users in our Twitter dataset, this computation is not practically feasible. Instead, our methodology for circumventing this problem was to sample  $10^9$  pairs of users uniformly at random, with replacement. We then compared the number of edges spanning a given distance to the estimated number of total user pairs for that same distance, given the sample.

### **8.3 Social Adoption of Hashtags and Future Users**

As we observed in the previous section, the topical affiliation structure of a social network is related to its social structure and can be useful to predict social links. In this section, we aim to exploit the information embedded in the structure of the network by using it to investigate its relation to the future popularity of the hashtags.

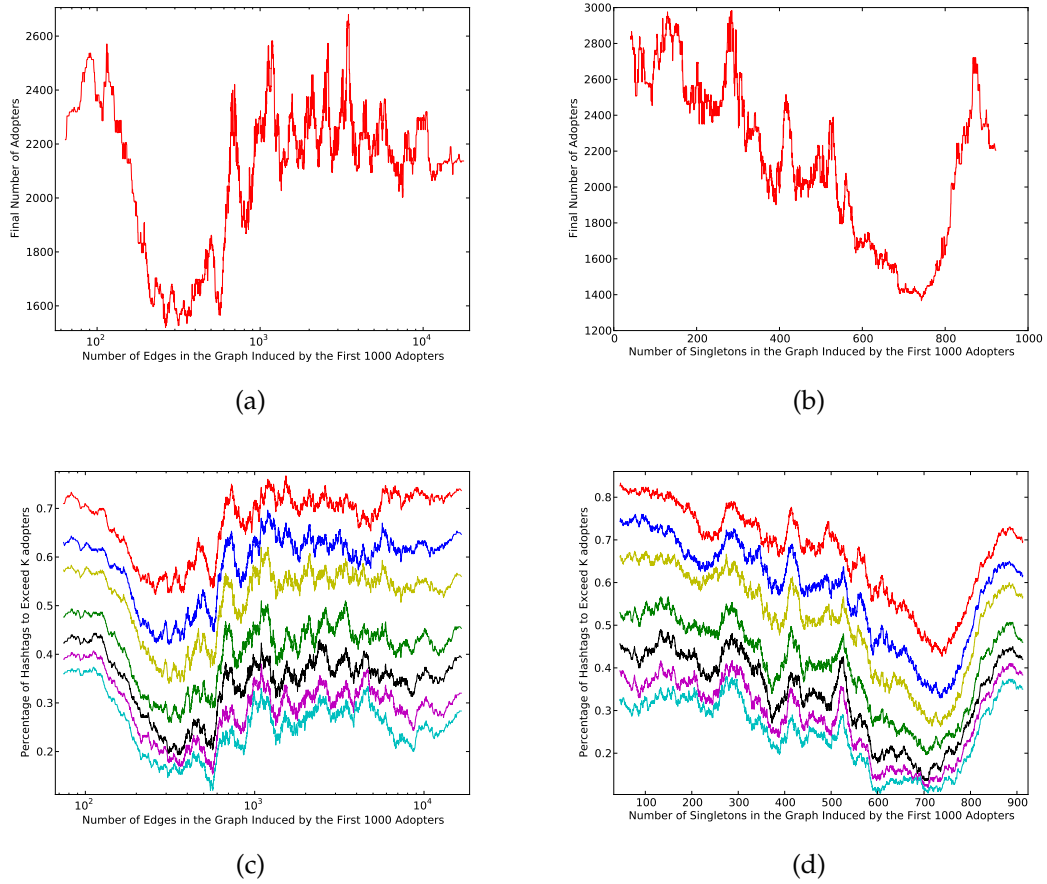


Figure 8.3: Median number of final adopters as a function of the number of (a) edges and (b) singletons in the graph induced by the 1000 initial adopters, using a sliding window. Probability that hashtags will exceed  $k$  adopters given the number of (c) edges and (d) singletons in the graph induced by the 1000 initial adopters, using a sliding window. From top to bottom,  $K = 1500, 1750, 2000, 2500, 3000, 3500, 4000$ . We observe that hashtags with many or few singletons and edges are more likely to grow than hashtags with intermediate amounts.

### 8.3.1 Correlations between initial graph structure and popularity

The properties of the graph of initial adopters of a hashtag tell us about the social structure of the initial adopters. This can in turn suggest properties of the



diffusion mechanism of the hashtag. For example, if the graph has a very large number of edges in it, it suggests that the users could have found out about the hashtag from each other and that many of the friends who haven't adopted it yet are likely to do so in the future. On the other hand, if there are very few edges in the graph, this suggests that the initial adopters did not discover the hashtag through their connections, since their connections had not adopted it yet, which means that users are not "virally" adopting the hashtag. It is an interesting question whether the eventual growth of the hashtag depends on the number of edges in the initial graph, and if so, whether large growth is correlated with a high or a low number of edges. Of course, more detailed properties of the graphs such as the number of connected components, the number of singletons, and the size of the largest component could also be important.

We begin by exploring how different structural properties of the initial graph affect the probability that a hashtag will grow. We consider all 7397 hashtags in our data that had at least 1000 adopters, and construct the follower graph induced by these 1000 users. For each hashtag, we look at the number of users that eventually used the hashtag and compute the number of edges and singletons in their corresponding initial graphs. Figures 8.3(a) and 8.3(b) show that that number of eventual adopters does not monotonically change with the number of singletons or edges, instead we find an interior minimum. This suggests that hashtags with either many or few edges and singletons tend to grow more than hashtags with an intermediate number of singletons and edges.

In practice, often times one is not interested in the exact final number of adopters in a diffusion process, instead it is desirable to know if the number of adopters will surpass a certain threshold or if the adopter population will

double, triple, etc. In Figures 8.3(c) and 8.3(d) we plot the probability that a hashtag with 1000 adopters will reach 1500, 1750, 2000, 2500, 3500, and 4000 adopters as a function of the number of singletons and edges in the subgraph of the initial 1000 adopters. As we found when we were asking about the final number of adopters, the likelihood of growth is highest when the singleton and edge counts are either very large or very small. Furthermore, the trends are consistent for the different choices of  $k$ , suggesting that the trend holds for the short, medium, and long terms. Note that we conducted the same experiment with different numbers of initial adopters (the 2000 initial adopters and the 4000 initial adopters), and we observe the same results.

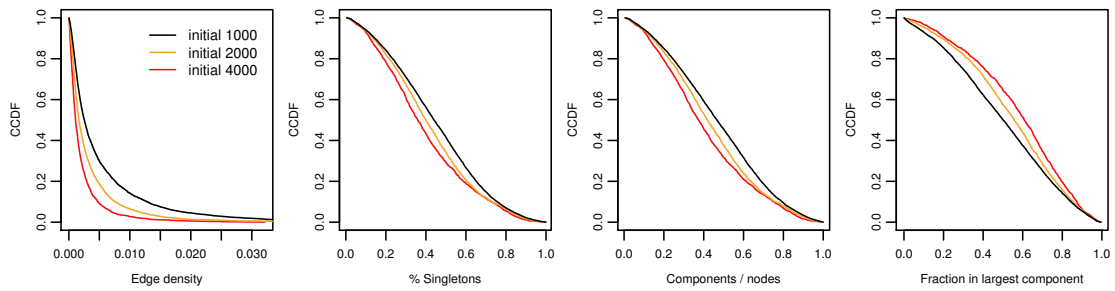


Figure 8.4: Distribution of the structural features of the subgraphs induced by the 1000, 2000, and 4000 initial adopters. We see that while the edge count exhibits a heavy tailed distribution, the number of singletons, components and the size of the largest component are all broadly distributed over their support.

### 8.3.2 Exogenous forces vs. Virality

We observe that hashtags that exhibit large growth tend to be those that either have initial adopters with a large number of connections among each other, or very few connections among each other. However, those hashtags that have a

medium number of connections among initial adopters do not tend to grow as much. While we have not conducted experiments that try to find evidence for any theories that explain this phenomenon, we discuss possible explanations.

Let us first consider why a small number of connections among initial users could imply large growth of the hashtag. Many of the hashtags our data set correspond to very popular world events or topics. For example *#iranelection* was used during the disputed 2009 elections, *#michaeljackson* was used around the time of the death of Michael Jackson, and *#iphone* is used around the time a new version of the iPhone is released. If we think about the initial adopters of these kinds of hashtags, it would not be surprising to encounter very few connections among them. What these hashtags have in common is a force exogenous to Twitter, making people adopt the hashtags independently of their Twitter connectivity. For example, the first set of people that used *#michaeljackson* could be the people that first found out about his death on the news and they were probably not friends on Twitter. Hence, a possible explanation for the fact that hashtags with sparse initial graphs tend to grow is that their initial graphs are sparse because the initial adopters are using the hashtag because of an exogenous force, and that exogenous force will also be responsible for the large growth of the hashtag.

On the other hand, we observed that hashtags with very dense initial graphs also tend to grow large. This can be explained by the “virality” argument: if the initial set of users are very well connected among each other, it must be that the hashtag is very sticky. That is, once a user gets exposed to the hashtag after seeing that her friend used it, she is very likely to use it. Then her friends that still haven’t adopted her will follow her example, and so on. Hence, the dense

Model Features	F: 1k→2k	F: 2k→4k	F: 4k → 8k	@≥3: 1k→2k	@≥3: 2k→4k	@≥3: 4k→8k
All	<b>0.674</b>	<b>0.668</b>	<b>0.683</b>	<b>0.568</b>	0.574	0.590
Social Graph Only	0.639	0.639	0.642	0.565	0.568	0.590
Full Graph Only	0.658	0.659	0.658	0.568	0.571	0.581
Info. Graph Only	0.642	0.649	0.663	0.559	<b>0.576</b>	<b>0.601</b>
# Full Graph Edges	0.556	0.525	0.548	0.534	0.506	0.547
# Social Edges	0.538	0.529	0.545	0.530	0.510	0.547
# Conn. Comps, Full Graph	0.579	0.591	0.603	0.525	0.511	0.535
# Conn. Comps, Social Graph	0.567	0.575	0.593	0.527	0.504	0.533
# Conn. Comps, Info. Graph	0.566	0.565	0.585	0.528	0.504	0.546
# Singletons, Full Graph	0.583	0.599	0.602	0.529	0.513	0.548
# Singletons, Social Graph	0.574	0.579	0.595	0.525	0.516	0.548
# Singletons, Info. Graph	0.571	0.566	0.590	0.525	0.513	0.545
Max Comp. Size, Full Graph	0.573	0.587	0.599	0.521	0.520	0.541
Max Comp. Size, Social Graph	0.555	0.581	0.598	0.520	0.520	0.537
Max Comp. Size, Info. Graph	0.556	0.565	0.586	0.527	0.504	0.547
Majority Vote	0.518	0.521	0.547	0.518	0.521	0.547

Table 8.2: Accuracy of a logistic regression mode for predicting whether a hashtag will double the number of adopters at different starting points: the 1000, 2000, and 4000 initial adopters, for both the follower and the @-message graphs. The accuracy of all models was evaluated using 10-fold cross-validation.

initial graph of a hashtag may signal that the hashtag is “going viral” and that explains why it will eventually obtain many adopters.

The hashtags that are in the middle of these two extremes lack both virality and exogenous force and hence do not obtain many adopters. Interestingly, we find that these two competing effects generate an interior minimum that we can observe at large scale in our data. However, we note that these theories are only meant as possible explanations and it is an open problem to design experiments that provide further evidence that these are the mechanism that explain the observed interior minimum.

### 8.3.3 Predicting growth from structure

We have seen how the number of singletons and edges of the initial graph can be informative with respect to a hashtag's growth. Now we would like to study of these two features, and more generally, the structure of the initial graph can actually predict whether the hashtag will obtain many additional adopters. In order to select appropriate features for a prediction model, it is important to understand the different kinds of connections we can differentiate on Twitter based on the following graph.

**Informational and social edges.** In Twitter, users can unilaterally follow or @-message other users without having to ask for their approval. This environment allows for the connections of users to have different meanings. For example, if two users on Twitter are friends in real life, they are likely to follow each other. However, if a user is interested in another user, but they do not actually know each other, we would expect the following relation may be unreciprocated. We refer to this type of relationship as informational. For example, celebrities are followed by many of their fans, but they don't usually follow their fans back. Hence, we could think of Twitter as a network composed in two kinds of edges: social and informational [25]. Given this lack of consistency of the meaning of connections on Twitter, we try to tease apart the two kinds of connections as they may have different prediction potential.

Formally, given the directed follower network on Twitter which we refer to as the *full graph*, we define an undirected edge between users  $A$  and  $B$  as *informational* if either  $A$  follows  $B$ , or  $B$  follows  $A$ , but not both. We define an undirected edge between users  $A$  and  $B$  as *social* if they follow each other. Next, we define the *social graph* as the network of Twitter users and their social edges

only, and the *informational graph* as the users with their informational edges only. Note that the social graph is undirected, and the full and informational graphs are directed. Note that we can define corresponding graphs using @-message edges instead of the follower edges in a similar way.

**The predictive model.** In this section, we train a logistic regression model to predict whether the number of adopters of a hashtag will eventually double. We examine other levels of growth further on. We use simple topological properties of the full graph, the social graph, and the informational graph. For each graph we compute the number of edges, number of singletons, number of connected components (weakly for the full and informational graphs), and the size of the largest connected component. We train separate logistic regression models with the same features but based on the @-message and follower graphs.

To understand our ability to meaningfully separate graphs based on the structural features we analyze, in Figure 8.3.1 we plot the complementary cumulative density functions for the features, as computed for all hashtags that exceeded size 1000, 2000, and 4000. We see the number of singletons, the number of components, and the number of adoptees in the largest component, the features are broadly distributed across their support, consistently for all three subgraph sizes.

For each feature, we include its value as well as the logarithm of the value. Additionally, for every feature except for number of edges, we include a “distance from the mid-point” transformation:  $|v_f - \frac{m_f}{2}|$ , where  $v_f$  is the value of the feature and  $m_f$  is the largest possible value of the feature. The reason for this transformation is that, as figure 8.3 suggests, high growth of the hashtags may be correlated with large or small values of some features. Having this trans-

formation allows the algorithm to capture this trend. We do not include this transformation for number of edges because for these features  $m_v$  is extremely large, and none of the hashtags we consider have an edge density greater than 0.5, making all these transformed features linearly dependent upon the initial feature.

We begin by using the logistic regression model to predict whether a hashtag will double in size. For each hashtag  $h$  that has at least  $k$  adopters, we compute the features of the model and predict whether it will eventually obtained  $2k$  adopters. We run the prediction task using the follower graph and the @-message graph separately. For the @-message graph we used a threshold of at least 3 @-messages to form an edge, having observed that different message count thresholds produced similar results. Table 8.3.3 shows the accuracy of the full multivariate model using all the features and transformations as well as a model using each standalone feature and its transformations. We evaluated the accuracy using 10-fold cross-validation for  $k = 1000, 2000, 4000$ . Using the follower graph we obtain an accuracy of around 67%. This is 14 percentage points above a baseline of around 53% obtained from a naive majority vote algorithm, which simply classifies all hashtags as “yes” if the majority of hashtags doubled and “no” if the majority of hashtags do not. Using the @-message graph we do not perform as well as with the follower graph. For the @-graph, the accuracy of the full model is 57% compared to a baseline of 53%. Furthermore, we find that the accuracy changes very little for the different choices of  $k$ , which suggests that classifying hashtags doubling does not get harder or easier we change the original size of the hashtag.

Lastly, we compare the performance under different sets of features, i.e., us-

ing the *Social Graph Only*, the *Full Graph Only* or the *Informational Graph Only*. It is shown that *Social Graph Only* cannot provide as good performance as the other two feature sets. The reason might be that informational relationships play a stronger role in the spread of hashtags. Furthermore, the *Informational Graph Only* performs marginally better than the social graph, and in some case it performs marginally better than the full model with all the features included. It is an interesting open problem to investigate if information edges indeed carry more predictive power than other kinds of edges, and if so, to determine possible explanations for it.

### 8.3.4 Longer prediction horizons

Having found that the original size of the hashtag does not affect the accuracy of the algorithm, we now investigate whether the accuracy changes as we change the horizon of prediction. That is, what happens to the accuracy if we try to predict whether the hashtag will grow by a factor of  $p$  for  $p \in (1, \infty)$ ? We expect that when  $p$  is close to 1 the algorithm will not gain much accuracy above the baseline for two reasons. First, the outcome will be very sensitive to noise since we are asking whether the hashtag will obtain just a few additional adopters, so finding the few hashtags that do not surpass the threshold becomes difficult. Second, because most hashtags will surpass the threshold, even the naive majority vote classifier will have high accuracy, leaving little room for improvement. Similarly, when  $p$  is very large, we expect that the structure of the graph will lose predictive power, as the graph is itself changing as additional users adopt the hashtag, and this evolution may change the nature of its growth. Also, since most hashtags will not surpass the threshold, the majority vote classifier will



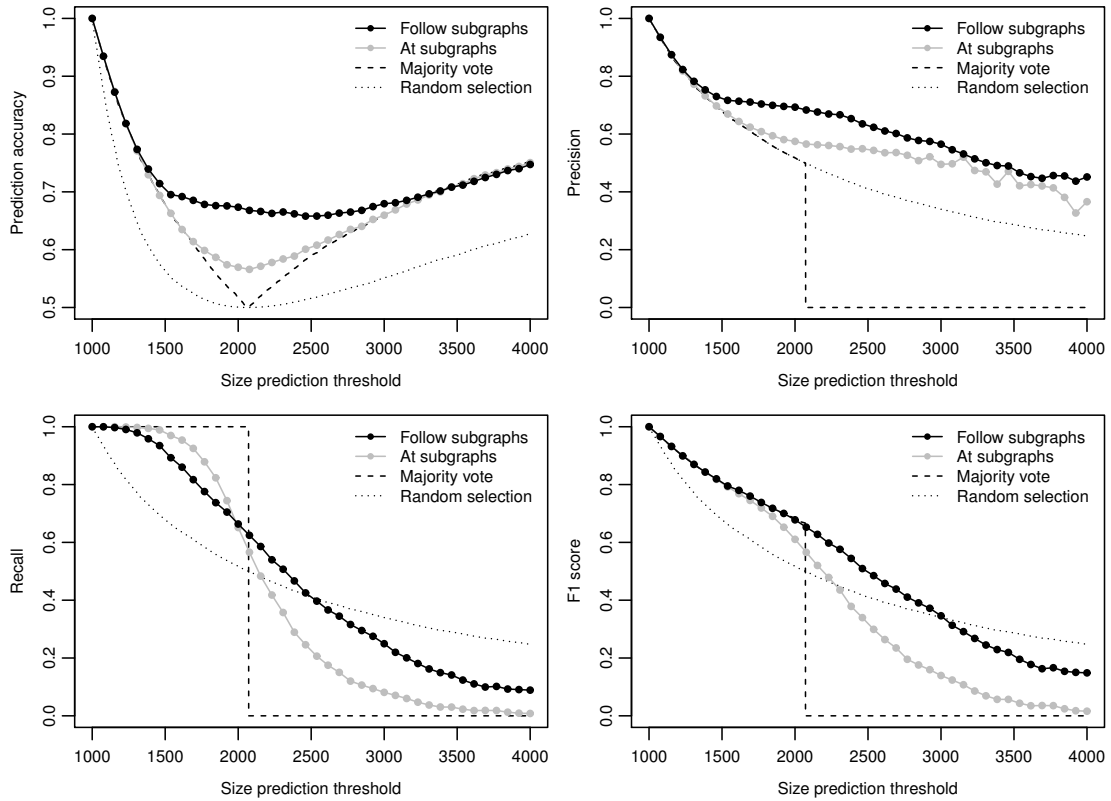


Figure 8.5: Prediction accuracy, precision, recall, and F1 score when predicting whether a hashtag will exceed a certain size using our logistic regression model based on graph structure. Models were trained using 5-fold cross validation, applied to those 7397 hashtags that reached a size of 1000.

again have high accuracy, again leaving little room for improvement.

To answer this question, we run logistic regression with the same features as above using the hashtags that had at least 1000 adopters and predicted whether they would reach at least  $M$  users. Figure 8.5 shows the accuracy, precision, recall, and F1 score of the full logistic model as a function of  $M$ . We compare these scores with two baseline naive classifiers – the majority vote classifier discussed above, and a random algorithm that classifies as “yes” a random set of hashtags of size equal to the fraction of hashtags that obtained at least  $M$  adopters.

We find that, indeed, the accuracy of our classifier is not above the baseline for large and small values of  $M$ . However, when we look at the precision of our classifier, we see that it stays above the baselines even for large values of  $M$ . That means that even for long term horizon prediction, the structure of the initial graphs maintains predictive power. The recall of the classifier starts off reasonably large for small values of  $M$ , but drops as  $M$  gets very large. This is due to the fact that when  $M$  is large, the number of hashtags that surpass  $M$  adopters will be very small, making it very hard to identify all the few that did.

In summary, our classifier maintains an accuracy of roughly 70%. Its accuracy stays above our baselines for mid-term horizons and it equals the baselines for long and short term horizons. Its precision decreases slightly as the horizon increases, but it always stays above our baselines. Its recall starts out high, but it drops dramatically as the horizon increases. It is optimal for a classifier to have high precision and recall, and in our case we are able to maintain high precision, but recall falls for long term horizons. Depending on the actual application of the classifier, sometimes it may be more desirable to have high recall than precision and vice-versa. A possible applications for wanting to know the future size of a certain hashtag is market forecasting. If people are using hashtags that correspond to new products, and an analyst would like to use social media analysis to track which product to apply their advertising budget towards, having high precision is preferable over having good recall. On the other hand, low recall simply means that many products that became popular were not identified by the algorithm, so opportunities were lost, but not investments.

## 8.4 Related Work

Collaborative tagging systems, which allow users to share their tags for particular resources, form the basis of our understanding of hashtags [93, 106, 56, 122, 132, 50, 117]. Marlow et al. [93] performed an early study on Flickr, developing a taxonomy of social tagging systems. They found that friends show a larger similarity in vocabulary compared with a random user baseline, which suggested that social links and tag usage is indeed related. Ramage et al. [106] proposed a generative model based on Latent Dirichlet Allocation (LDA) that jointly models text and tags in such settings. Markines et al. [91] employed similarity measures such as matching, overlap, mutual information, and Jaccard, Dice, and cosine similarity to study topics. Recent work by Schifanella et al. [117] also works on the interplay of the social and semantic components of social media on Flickr and Last.fm. They showed that a substantial level of local lexical and topical alignment is observable among users who lie close to each other in the social network. Their analysis suggests that users with similar topical interests are more likely to be friends, and therefore semantic similarity measures among users based solely on their annotation metadata should be predictive of social links.

## 8.5 Discussion

In this study, we find that the user-generated hashtag set system on Twitter and the topology of the connections among users are two fundamentally related structures. Furthermore, as proof of concept, we show that our findings could be useful for predicting social links and popularity even when using simple fea-

tures and a standard predictive model. By studying distance measures among users, based on the topical proximity embedded by the hashtag set system, we are able to predict, with reasonably high accuracy, the links between the users. Furthermore, the size of the smallest common hashtag turns out to be a very good predictor of linkage despite being one of the cheapest ones to compute. We also found that combining the topical overlap of the users with the structural features of the graph induced by the shared topics can dramatically improve the accuracy of the prediction task. A possible application of this would be situations where one knows the topical interests of users and would like to predict connections among them, including recommendations systems for connections in online social networks. For example, recommending who to follow on Twitter based on an individual's hashtag usage, or recommending people to friend on Facebook based on an individual's "likes."

After observing how simple structural features of the graph induced by a hashtag were useful in predicting social connections, we discovered that they are also useful for predicting the popularity of the hashtag itself. These features are very efficient to compute in  $O(|V| + |E|)$  time. We found the future popularity of the hashtags does not monotonically increase with the density of the graph induced by its initial set of users. Instead, we find that the popularity of the hashtag is highest when the density is either very low or very high. This is an interesting finding that offers a different perspective to the ideas from "viral marketing" where a small number of connections could be considered a negative property with respect to future growth.

Throughout our study we compare our results generated by the follower graph and the @-message graph. We find interesting distinctions among the

two. For example, @-connections are easier to predict, but they turn out to be less informative when predicting the popularity of a hashtag from the connections of early adopters. Also, since the @-graph can be viewed as a weighted graph, we are able compare our results on different levels of edge strength by considering graph with only @-edges with at least  $k$  @-messages. We find that stronger ties are easier to predict, but they do not provide better or worse predictive power with regard to future hashtag popularity.

## CHAPTER 9

### CONCLUSION

The study of social networks has been an important area of research in sociology for a long time. Through their work, we have acquired a vast amount of knowledge and intuition about the dynamics of social networks. With the advent of the Web comes a unique opportunity to study social networks and information flow from the perspective of computer science and mathematics. A growing body of research on social networks from these fields has emerged in recent years. Data from online social networks, together with the knowledge we already have from sociology, has been used to validate known theories, propose new theories, and conduct empirical studies at a very large scale, which would have been nearly impossible to do before the Web. New algorithms and mathematical models continue to advance our understanding of social networks and the complex processes that occur on them. Throughout this work, we use large data sets, mathematical models, and algorithms to analyze two major aspects of online social networks: how they are formed and evolve over time, and how information flows through them.

Online social networks are a good proxy for studying social relationships since many people who know each other offline interact online. However, there are important differences between social networks online and offline. In chapter 3, we discuss one of those differences. In most online social networks, connections between nodes do not always signal social relationships. Sometimes online edges signal that a person is interested in the information they can obtain from the other, and not that the two actually know each other offline. This is especially true in directed online social networks such as Twitter where fans

can follow celebrities who do not have to follow their fans back. This difference implies that Twitter is a combination of two kinds of networks, which have been studied separately, but not when they are combined in a single domain: social networks and information networks.

In chapter 4, we begin to study the formation of social-information networks and we find a mechanism that is well-understood for social networks but not for social-information networks. We propose a generalization of the well-known principle of triadic closure for directed social-information networks, which we call *directed closure*. Through a randomization test we find evidence of the significant presence of directed closure in the formation of the Twitter network. Furthermore, we find that there is great variation in the amount of directed closure among the users who follow different celebrities. Through a series of network generating models, we find explanations for this variation.

Besides studying the formation of social networks, it is important to understand how the interaction among connected people changes as a response to new edges arriving to the network. This is particularly important when we are interested in studying information diffusion on social networks. In chapter 5, we analyze how communication among nodes involved in open triads changes after the triad closes. We compare our findings with the predictions that sociological theories would suggest. Overall, we find that our findings are consistent with balance theory, which suggests that a closed triad should be in some sense stronger than an open one, and hence communication among the nodes should increase after the triad closes.

One important aspect of social networks is that they control, in part, the spread of information among people. A large body of research that studies in-

formation flow through data from social media sites has emerged in the past few years. We have gained a tremendous understanding of the dynamics of information diffusion on social networks, which generalize to many different kinds of information. Because we now have access to very rich data sets that provide information, not only about who said something when, but also about the content of what was said, we are able to study information flow in much greater detail. In chapter 6, we study the mechanics of information flow that operate differently depending on the topic of the information. We are able to validate previous sociological theories at large scale. For example, we find that politically controversial topics tend to spread more when people on the networks are exposed to them many times. On the other hand, non-controversial topics do not benefit as much as controversial ones from nodes being exposed to them multiple times. This is consistent with the complex contagion principle from sociology.

Much of the research on information diffusion, including the work in chapter 6, makes the simplifying assumption that when a user is exposed to information from  $k$  other users, each one of those  $k$  users have the same effect on the exposed user. In other words, the assumption is that all users exert the same amount of influence on their neighbors. This assumption is often made because it allows us to study simple models of information diffusion. However, enhancing models by allowing nodes to have different amount of influence is a natural future direction for information diffusion research. In order to incorporate influence in the study of information diffusion, it is necessary to build tools that allows us to define and measure influence. In chapter 7, we show that studying information flow on social networks can be useful for measuring the influence that users have on the network. We propose an algorithm that ranks nodes based



on their influence in spreading information in the network and making nodes that are not usually active in the network become active. We validate our approach by testing the extent to which it can predict the traffic that URLs posted by influential Twitter users will receive.

Information flow and social network evolution are intimately connected to each other. In chapter 8, we show that the structure of the Twitter social network is related to the structure of the network of topical affiliations of Twitter users. We find that the structure of the social network can predict the future popularity of the topics that users will post about, and that the topics that users post about can be used to predict new links in the network. Furthermore, we find that the prediction of links and popularity can yield reasonably high accuracies even when using very simple and computational inexpensive measure of the structure of the social and topical affiliation networks.

While much progress has been made in advancing the understanding of social networks through the analysis of data from social media sites, much of the data still remains to be studied. In particular, most of the actual content generated by users of social media sites has not been looked at as much as the connections among the users. In chapter 6 and 8, we begin to study topics in relation to information flow and link formation, but only through hashtags or labels of the messages posted. It remains as future work to include other features of the content as part of our analysis. For example, in the case of political topics, one interesting feature to study would be how the tone or sentiment affects the spread of political ideas. It is challenging to answer very detailed questions like this one at large scale. However, it is likely that many important features of the dynamics of information flow hide under the more subtle aspects

of the information and not just under its topic.

Another aspect of this research that remains as future work is measuring spread of behaviors on social networks based on more costly decisions by the nodes. In chapter 6, we measured the probability that a user would tweet about a political topic after  $k$  friends had done so. In chapter 7, we measured influence by the extent to which users are able to get their messages forwarded by other users. In both cases, the action was simply to post a short message in a social media site. Measuring influence and behavior spread based on actions that require more effort such as buying a product, stopping smoking, or joining a protest in person, may yield different results. It would be interesting to study the spread of more complex contagions at large scale.

Finally, as we have seen in chapters 3 and 4, online social networks are a mixture of social and information networks. We can use proxies for identifying which edges are social and which are informational, but differentiating between social edges and informational edges with high confidence remains as a task for future work. An interesting experiment would be to join an offline social network with an online one to study the differences between the two in more depth. This could shed light into the structural differences between the two. Furthermore, understanding the role that informational ties play on information flow online could yield important results about the differences between online and offline information diffusion.

## BIBLIOGRAPHY

- [1] L.A. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.
- [2] Lada Adamic and Natalie Glance. The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, pages 36–43, 2005.
- [3] Eytan Adar, Li Zhang, Lada A. Adamic, and Rajan M. Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem*, 2004.
- [4] Nitin Agarwal, Huan Liu, Lei Tang, and Philip S. Yu. Identifying the influential bloggers in a community. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, pages 207–218, New York, NY, USA, 2008. ACM.
- [5] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.
- [6] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. Influence and correlation in social networks. In *Proc. 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 7–15, 2008.
- [7] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci. USA*, 106(51):21544–21549, December 2009.
- [8] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: Membership, growth, and evolution. In *Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.
- [9] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone’s an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM.
- [10] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

- [11] Ginestra Bianconi and Albert-László Barabási. Bose-Einstein condensation in complex networks. *Physical Review Letters*, 86:5632–5635, 2001.
- [12] D. Boyd, S. Golder, and G. Lotan. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, volume 0, pages 1–10, Los Alamitos, CA, USA, January 2010. IEEE.
- [13] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. In *Proc. 7th International World Wide Web Conference*, pages 107–117, 1998.
- [14] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the Web. In *Proc. 9th International World Wide Web Conference*, pages 309–320, 2000.
- [15] Jacqueline Johnson Brown and Peter H Reingen. Social ties and word-of-mouth referral behavior. *Journal of Consumer Research*, 14(3):350–62, 1987.
- [16] Michael J. Brzozowski, Tad Hogg, and Gábor Szabó. Friends and foes: ideological social networking. In *Proc. 26th ACM Conference on Human Factors in Computing Systems*, pages 817–820, 2008.
- [17] Michael J. Brzozowski and Daniel M. Romero. Who should i follow? recommending people in directed social networks. In *ICWSM*, 2011.
- [18] Ronald S. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, 1992.
- [19] Ronald S. Burt. Structural holes and good ideas. *American Journal of Sociology*, 110(2):349–99, September 2004.
- [20] Dorwin Cartwright and Frank Harary. Structure balance: A generalization of Heider’s theory. *Psychological Review*, 63(5):277–293, September 1956.
- [21] Damon Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, 3 September 2010.
- [22] Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American Journal of Sociology*, 113:702–734, 2007.

- [23] Meeyoung Cha, Hamed Haddadi, Fabrcio Benevenuto, and Krishna P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *in ICWSM 10: Proceedings of international AAAI Conference on Weblogs and Social*, 2010.
- [24] Jilin Chen, Werner Geyer, Casey Dugan, Michael Muller, and Ido Guy. Make new friends, but keep the old: recommending people on social networking sites. In *Proceedings of the 27th international conference on Human factors in computing systems, CHI '09*, pages 201–210, New York, NY, USA, 2009. ACM.
- [25] Justin Cheng, Daniel Romero, Brendan Meeder, and Jon Kleinberg. Predicting reciprocity in social networks. In *ICSC*, 2011.
- [26] Dan Cosley, Daniel P. Huttenlocher, Jon M. Kleinberg, Xiangyang Lan, and Siddharth Suri. Sequential influence models in social networks. In *Proc. 4th International Conference on Weblogs and Social Media*, 2010.
- [27] David Crandall, Dan Cosley, Dan Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. In *Proc. 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 160–168, 2008.
- [28] Riley Crane and Didier Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proc. Natl. Acad. Sci. USA*, 105(41):15649–15653, 29 September 2008.
- [29] James A. Davis. Clustering and structural balance in graphs. *Human Relations*, 20(2):181–187, 1967.
- [30] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proc. 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 57–66, 2001.
- [31] R I M Dunbar. Neocortex size and group-size in primates - a test of the hypothesis. *Journal of Human Evolution*, 28(3):287–296, 1995.
- [32] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.
- [33] Richard M. Emerson. Power-dependence relations. *American Sociological Review*, 27:31–40, 1962.

- [34] Wojciech Galuba, Dipanjan Chakraborty, Karl Aberer, Zoran Despotovic, and Wolfgang Kellerer. Outtweeting the Twitterers - Predicting Information Cascades in Microblogs. In *Proceedings of the 3rd Workshop on Online Social Networks (WOSN 2010)*, 2010.
- [35] Eric Gilbert and Karrie Karahalios. Predicting tie strength with social media. In *Proceedings of the 27th international conference on Human factors in computing systems, CHI '09*, pages 211–220, New York, NY, USA, 2009. ACM.
- [36] Eric Gilbert and Karrie Karahalios. Predicting tie strength with social media. In *Proc. 27th ACM Conference on Human Factors in Computing Systems*, pages 211–220, 2009.
- [37] Eric Gilbert, Karrie Karahalios, and Christian Sandvig. The network in the garden: Designing social media for rural life. *American Behavioral Scientist*, 53(9):1367–1388, 2010.
- [38] Malcolm Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference*. Little, Brown, 2000.
- [39] Malcolm Gladwell. Small change: Why the revolution will not be tweeted. *The New Yorker*, 4 October 2010.
- [40] J. Goldenberg, B. Libai, and Muller. Using complex systems analysis to advance marketing theory development. *Academy of Marketing Science Review*, 2001.
- [41] Scott A. Golder, Dennis Wilkinson, and Bernardo A. Huberman. Rhythms of social interaction: Messaging within a massive online network. In *Proc. 3rd International Conference on Communities and Technologies*, 2007.
- [42] Scott A. Golder and Sarita Yardi. Structural predictors of tie formation in twitter: Transitivity and mutuality. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing, SOCIALCOM '10*, pages 88–95, Washington, DC, USA, 2010. IEEE Computer Society.
- [43] Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining, WSDM '10*, pages 241–250, New York, NY, USA, 2010. ACM.

- [44] Mark Granovetter. The strength of weak ties. *American Journal of Sociology*, 78:1360–1380, 1973.
- [45] Mark Granovetter. *Getting a Job: A Study of Contacts and Careers*. University of Chicago Press, 1974.
- [46] Mark Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83:1420–1443, 1978.
- [47] Mark Granovetter. The strength of weak ties: A network theory revisited. *Sociological Theory*, 1:201–233, 1983.
- [48] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *WWW*, 2004.
- [49] Daniel Gruhl, David Liben-Nowell, R. V. Guha, and Andrew Tomkins. Information diffusion through blogspace. In *Proc. 13th International World Wide Web Conference*, 2004.
- [50] Harry Halpin, Valentin Robu, and Hana Shepherd. The complex dynamics of collaborative tagging. In *WWW*, 2007.
- [51] Tony Hammond, Timo Hannay, Ben Lund, and Joanna Scott. Social bookmarking tools (i): A general review. *D-Lib Magazine*, 11, 2005.
- [52] Chip Heath and Dan Heath. *Made to Stick: Why Some Ideas Survive and Others Die*. Random House, 2007.
- [53] Fritz Heider. Attitudes and cognitive organization. *Journal of Psychology*, 21:107–112, 1946.
- [54] Fritz Heider. *The Psychology of Interpersonal Relations*. John Wiley & Sons, 1958.
- [55] J E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572, 2005.
- [56] Ming-Hung Hsu, Yu-Hui Chang, and Hsin-Hsi Chen. Temporal correlation between social tags and emerging long-term trend detection. In *ICWSM*, 2010.

- [57] Jeff Huang, Katherine M. Thornton, and Efthimis N. Efthimiadis. Conversational tagging in twitter. In *HT*, 2010.
- [58] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Crowdsourcing, attention and productivity. *Journal of Information Science*, 35(6):758–765, December 2009.
- [59] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1), January 2009.
- [60] Matthew O. Jackson. *Social and Economic Networks*. Princeton University Press, August 2008.
- [61] Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2169–2188, November 2009.
- [62] Anders Johansen. Probing human response times. *Physica A*, 338(1–2):286–291, 2004.
- [63] John Kalucki. Twitter, please explain how cursors work, 4 October 2009. [http://groups.google.com/group/twitter-development-talk/browse\\_thread/thread/c290d69c80cebb42](http://groups.google.com/group/twitter-development-talk/browse_thread/thread/c290d69c80cebb42).
- [64] Denise B. Kandel. Homophily, selection, and socialization in adolescent friendships. *American Journal of Sociology*, 84(2):427–436, September 1978.
- [65] David Kempe, Jon Kleinberg, and va Tardos. Influential nodes in a diffusion model for social networks. In *IN ICALP*, pages 1127–1138. Springer Verlag, 2005.
- [66] Masahiro Kimura and Kazumi Saito. Tractable models for information diffusion in social networks. In *PKDD*, 2006.
- [67] J. Kleinberg. Small-world phenomena and the dynamics of information. In *NIPS*, 2002.
- [68] Jon Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999. A preliminary version appears in the Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, Jan. 1998.



- [69] Jon Kleinberg. The small-world phenomenon: an algorithmic perspective. In *Proc. 32nd ACM Symposium on Theory of Computing*, pages 163–170, 2000.
- [70] Jon Kleinberg. The convergence of social and technological networks. *Communications of the ACM*, 51(11):66–72, 2008.
- [71] Gueorgi Kossinets and Duncan Watts. Empirical analysis of an evolving social network. *Science*, 311:88–90, 2006.
- [72] Gueorgi Kossinets and Duncan Watts. Origins of homophily in an evolving social network. *American Journal of Sociology*, 115(2):405–50, September 2009.
- [73] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *KDD*, 2006.
- [74] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew Tomkins, and Eli Upfal. Stochastic models for the web graph. In *Proc. 41st IEEE Symposium on Foundations of Computer Science*, pages 57–65, 2000.
- [75] Jérôme Kunegis, Andreas Lommatzsch, and Christian Bauckhage. The slashdot zoo: mining a social network with negative edges. In *Proceedings of the 18th international conference on World wide web, WWW '09*, pages 741–750, New York, NY, USA, 2009. ACM.
- [76] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web, WWW '10*, pages 591–600, New York, NY, USA, 2010. ACM.
- [77] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *WWW*, 2010.
- [78] Paul Lazarsfeld and Robert K. Merton. Friendship as a social process: A substantive and methodological analysis. In Morroe Berger, Theodore Abel, and Charles H. Page, editors, *Freedom and Control in Modern Society*, pages 18–66. Van Nostrand, 1954.
- [79] Jure Leskovec, Lada Adamic, and Bernardo Huberman. The dynamics of viral marketing. In *Proc. 7th ACM Conference on Electronic Commerce*, 2006.

- [80] Jure Leskovec, Lada Adamic, and Bernardo Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 1(1), May 2007.
- [81] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 2007.
- [82] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009.
- [83] Jure Leskovec, Dan Huttenlocher, and Jon Kleinberg. Predicting positive and negative links in on-line social networks. In *Proc. 19th International World Wide Web Conference*, pages 641–650, 2010.
- [84] Jure Leskovec, Dan Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *Proc. 28th ACM Conference on Human Factors in Computing Systems*, pages 1361–1370, 2010.
- [85] Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. Statistical properties of community structure in large social and information networks. In *WWW*, 2008.
- [86] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. Cascading behavior in large blog graphs. In *Proc. SIAM International Conference on Data Mining*, 2007.
- [87] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management, CIKM '03*, pages 556–559, New York, NY, USA, 2003. ACM.
- [88] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, 58:1019–1031, May 2007.
- [89] David Liben-Nowell and Jon Kleinberg. Tracing information flow on a global scale using Internet chain-letter data. *Proc. Natl. Acad. Sci. USA*, 105(12):4633–4638, March 2008.
- [90] R. Dean Malmgren, Daniel B. Stouffer, Adilson E. Motter, and Luis A. N. Amaral. A poissonian explanation for heavy tails in e-mail communication. *Proc. Natl. Acad. Sci. USA*, 105(47):18153–18158, 25 November 2008.

- [91] Benjamin Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, and Stum Gerd. Evaluating similarity measures for emergent semantics of social tagging. In *WWW*, 2009.
- [92] Cameron Marlow, Lee Byron, Tom Lento, and Itamar Rosenn. Maintained relationships on Facebook, 2009. On-line at <http://overstated.net/2009/03/09/maintained-relationships-on-facebook>.
- [93] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HT*, 2006.
- [94] Peter V. Marsden and Karen E. Campbell. Measuring Tie Strength. *Social Forces*, 63(2):482–501, December 1984.
- [95] Doug McAdam. Recruitment to high-risk activism: The case of Freedom Summer. *American Journal of Sociology*, 92:64–90, 1986.
- [96] Doug McAdam. *Freedom Summer*. Oxford University Press, 1988.
- [97] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.
- [98] Filippo Menczer. Growing and navigating the small world Web by local content. *Proc. Natl. Acad. Sci. USA*, 99(22):14014–14019, October 2002.
- [99] Stanley Milgram. The small-world problem. *Psychology Today*, 2:60–67, 1967.
- [100] James Moody. Race, school integration, and friendship segregation in america. *American Journal of Sociology*, 107(3):679–716, November 2001.
- [101] M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.
- [102] Mark E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.
- [103] Liben D. Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *PNAS*, 102(33):11623–11628, 2005.

- [104] Rin Popescul and Lyle H. Ungar. Statistical relational learning for link prediction. In *In Proceedings of the Workshop on Learning Statistical Models from Relational Data at IJCAI-2003*, 2003.
- [105] Derek De Solla Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, pages 292–306, 1976.
- [106] Daniel Ramage, Paul Heymann, Christopher D. Manning, and Hector Garcia-Molina. Clustering the tagged web. In *WSDM*, 2009.
- [107] Anatole Rapoport. Spread of information through a population with socio-structural bias I: Assumption of transitivity. *Bulletin of Mathematical Biophysics*, 15(4):523–533, December 1953.
- [108] Ray Reagans. Preferences, identity, and competition: Predicting tie strength from demographic data. *Manage. Sci.*, 51(9):1374–1383, September 2005.
- [109] Matthew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD*, 2002.
- [110] Everett Rogers. *Diffusion of Innovations*. Free Press, fourth edition, 1995.
- [111] Everett M. Rogers. *Diffusion of Innovations, Fourth Edition*. Free Press, 1995.
- [112] Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. Influence and passivity in social media. In *ECML/PKDD (3)*, pages 18–33, 2011.
- [113] Daniel M. Romero and Jon Kleinberg. The Directed Closure Process in Hybrid Social-Information Networks, with an Analysis of Link Formation on Twitter. In *ICWSM*, 2010.
- [114] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter. In *WWW*, 2011.
- [115] Daniel M. Romero, Chenhao Tan, and Johan Ugander. Social-topical affiliations: The interplay between structure and popularity. *CoRR*, abs/1112.1115, 2011.

- [116] Daniel Mauricio Romero, Brendan Meeder, Vladimir Barash, and Jon M. Kleinberg. Maintaining ties on social media sites: The competing effects of balance, exchange, and betweenness. In *ICWSM*, 2011.
- [117] Rossano Schifanella, Alain Barrat, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Folks in folksonomies: social link prediction from shared metadata. In *WSDM*, 2010.
- [118] X. Shi, L. Adamic, and M. Strauss. Networks of strong ties. *Physica A: Statistical Mechanics and its Applications*, 378(1):33–47, May 2007.
- [119] David Strang and Sarah Soule. Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annual Review of Sociology*, 24:265–290, 1998.
- [120] Eric Sun, Itamar Rosenn, Cameron Marlow, and Thomas M. Lento. Gesundheit! Modeling contagion through Facebook News Feed. In *Proc. 3rd International Conference on Weblogs and Social Media*, 2009.
- [121] Ben Taskar, Ming fai Wong, Pieter Abbeel, and Daphne Koller. Link prediction in relational data. In *NIPS*, 2003.
- [122] Jennifer Thom-Santelli, Michael J. Muller, and David R. Millen. Social tagging roles: publishers, evangelists, leaders. In *CHI*, 2008.
- [123] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.
- [124] Alexei Vazquez. Knowing a network by walking on it: Emergence of scaling. Technical Report cond-mat/0006132, arxiv.org, June 2000.
- [125] Alexei Vazquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E*, 67(056104), 2003.
- [126] Alexei Vazquez, Joao Gama Oliveira, Zoltan Deszo, Kwang-Il Goh, Imre Kondor, and Albert-Laszlo Barabasi. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(036127), 2006.
- [127] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

- [128] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393:440–442, 1998.
- [129] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twiterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM ’10, pages 261–270, New York, NY, USA, 2010. ACM.
- [130] David Willer (editor). *Network Exchange Theory*. Praeger, 1999.
- [131] Fang Wu, Bernardo A. Huberman, Lada A. Adamic, and Joshua R. Tyler. Information flow in social groups. *Physica A: Statistical Mechanics and its Applications*, 337(12):327 – 335, 2004.
- [132] Zhijun Yin, Rui Li, Qiaozhu Mei, and Jiawei Han. Exploring social tagging graph for web object classification. In *KDD*, 2009.